View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by NORA - Norwegian Open Research Archives

Hvem er forfatteren? - Stilometriske undersøkelser av norske prosatekster

Victoria Troland

Masteroppgave i datalingvistikk og språkteknologi

Institutt for lingvistiske, litterære og estetiske studium
Universitetet i Bergen

2015

UNIVERSITETET I BERGEN

UNIVERSITETET I BERGEN

Institutt for lingvistiske, litterære og estetiske studium

DASP350
Masteroppgave i datalingvistikk og språkteknologi
Vårsemester 2015

Hvem er forfatteren? - Stilometriske undersøkelser av norske prosatekster

Victoria Troland

Sammendrag

Stilometri er kvantitative undersøkelser av skrivestil ved hjelp av lingvistiske trekk. Ved hjelp av stilometriske metoder kan forfatterskap, egenskaper til forfatteren og hvorvidt en tekst er skrevet av en eller flere forfattere fastslås. Tidligere er stilometri blitt brukt til å oppdage plagiering og identifisere forfattere av tekster hvor denne er ukjent.

Denne studien utforsker stilometriske metoder på norske prosatekster. Målet er å undersøke om det er mulig å fastslå forfatteren ved hjelp av forskjellige typer lingvistiske trekk og statistiske metoder. Trekkene som er hentet ut er ord- og tegnbaserte sammen med et primært syntaktisk trekksett. Enkelte av trekkene er særnorske. De statistiske metodene som er brukt er overvåkede og ikke-overvåkede metoder.

Resultatene viser at det er mulig fastslå forfattere av norske prosatekster. Trekksettet ekstrahert fra trebanken INESS oppnådde 76.2 % riktig klassifisering med maskinlæring. Resultatene av de leksikalske trekkene avdekket underliggende likheter av tekstene som indikerte samme forfatter. De leksikalske trekkene klarte ofte å bidra til å gruppere forfattere riktig.

Stylometry is the quantitative research of stylistic differences by way of linguistics features. It is used to determine authorship of a text, author profiling and to discover whether a text has one or more authors. Stylometric research has previously been used in tasks to discover plagiarizing, author verification and in authorship attribution of texts where authorship is unknown.

This study aims to apply stylometric methods to Norwegian prose in order to determine authorship attribution. The features used in this study were word- and character-based and predominately a syntactic feature set, with some features being language specific. The statistical methods ranged from supervised to unsupervised methods.

The results indicate that authorship attribution is possible on Norwegian prose. The syntactic features extracted from the treebank INESS resulted in a classification rate of 76.2 % at the highest of correctly classified instances. Lexical features, that were character- and word-based, were able to indicate textual closeness to suggest authorship. The results show that the features are able to correctly suggest authorship.

Forord

Dette prosjektet har vært en stor del av livet mitt det siste året. På veien har jeg fått gode råd, veiledning, hjelp og motivasjon. Jeg vil derfor benytte muligheten her til å takke alle som har vært delaktig og til hjelp i løpet prosessen.

Jeg vil først og fremst takke veilederen min, Prof. Koenraad De Smedt, som gjennom året har vært tilgjengelig for spørsmål og tatt seg tid til å veilede. Han har en egen evne til å alltid se det viktigste og mest essensielle, noe som har vært uvurderlig. Noen av tabellene og grafene i studien er basert på R-skript skrevet av Koenraad De Smedt eller modifisert ut fra hans skript.

En stor takk til Paul Meurer som har bistått med teknisk assistanse, både med Aviskorpus ann. og INESS. Uten hans vilje til å tilrettelegge teknisk og gi kjappe tilbakemeldinger ville studien ha blitt svært annerledes.

Alle korrekturleserne mine, mamma, Odd Ivar, Mariel og Ragnhild fortjener en stor takk. Med sine skarpe øyne og objektivitet hjalp de med å gjøre teksten leselig for andre enn meg selv.

Jeg vil også takke mine medstudenter som har har tatt del i prosessen med sin faglig entusiasme og støtte. Spesielt vil jeg takke datalingvistene Anja og Håvar, lingvistene Eli, Julie, Christina, Klaas og digital kultur-folkene på lesesalen.

Innhold

Kapittel 1

Innledning

1.1 Innledning og hypotese

I 2013 ble boken *The Cuckoo's Calling* under pseudonymet Robert Galbraith utgitt. En journalist i The Sunday Times ble tipset om at forfatteren bak boken var J. K. Rowling (Lyall, 2013). Til tross for at journalisten avdekket flere sammenhenger mellom Rowling og Galbraith, som at de hadde samme agent, redaktør og forlag, var dette ikke nok til å bevise at Rowling var forfatteren bak pseudonymet.

Journalisten sendte deretter boken til datalingvistiske eksperter for å utføre en stilometrisk[1] undersøkelse for å verifisere at Rowling var Robert Galbraith. De konkluderte med at boken var skrevet av Rowling.

Min problemstilling

I denne studien kommer jeg til å undersøke norske prosatekster og teste forskjellige stilometriske metoder for å fastslå forfatterskap. Hypotesen min er:

- Kan forfattere av prosatekster gjenkjennes ved hjelp av stilometriske metoder, og i så tilfelle hvilke metoder?

Metodene varierer i typer lingvistiske trekk og statistiske algoritmer. Tekstene ble hentet fra trebanken INESS [2] og verktøyene jeg har benyttet er *Stylo* (Eder et al., 2013) og WEKA (Hall et al., 2009).

1.1.1 Kort om stilometri

Stlometri forsøker å måle og analysere variasjonen innenfor skrivestiler, og prøver å måle likhetene og forskjellene mellom forfattere, sjangere, perioder, o.l. Stilometriske metoder kan blant annet brukes til forfatterverifisering, forfatterattribuering, å oppdage spam, plagiering og forfatterprofilering.

[1] Begrepet er oversatt av Koenraad De Smedt fra engelske *stylometry*

[2] `<http://clarino.uib.no/iness/page>`

Forfatterprofilering vil si å oppdage egenskaper til en forfatter ut ifra tekst. Egenskapene kan være kjønn, alder, humør og personlighet (Schler et al., 2006; Keshtkar and Inkpen, 2009; Luyckx and Daelemans, 2008b).

Analysene blir gjort ved at man identifiserer forskjellige trekk og analyserer dem statistisk for å gjenkjenne skrivestiler. Eksempler på trekk kan være ordfrekvens, setningslengde og ordklassebruk. Trekk og trekktyper vil jeg komme innpå etter hvert.

I nyere tid er metodene blitt mer varierte og avanserte, blant annet har maskinlæringsmodeller blitt inkludert. I tillegg er typene trekk som blir analysert blitt mer flerfoldige og inkluderer nå blant annet syntaktiske trekk.

1.2 Studiens oppbyggning

I kapittel 2 diskuteres teorien rundt stilometri. Jeg diskuterer forskjellige statistiske metoder, tidligere forskning, mulige lingvistiske trekk og stilometri på norsk. Dessuten tar jeg opp forutsetningene innen stilometri og kritikk av stilometrisk forskning.

Kapittel 3 beskriver metodene som er brukt i undersøkelsene. Der beskrives hvilke tekster korpuset består av, hvordan korpuset er valgt ut og en vurdering av korpuset. I tillegg beskrives utførelsen av forsøkene og programmene som er brukt som verktøy.

I kapittel 4 beskrives undersøkelsene med programmet *Stylo* og med forskjellige variasjoner av korpuset beskrevet i kapittel 3. Hensikten er å undersøke om tekster av samme forfatter kan kategoriseres sammen ved bruk av læringsalgoritmer. Resultatene blir til slutt diskutert.

I kapittel 5 beskrives undersøkelsene av korpustekstene og syntaktiske og morfo-syntaktiske trekk hentet fra trebanken INESS. Frekvensene av trekkene er modellert med forskjellige maskinlæringsalgoritmer for å undersøke om forfattere kan identifiseres ut i fra modellene. Deretter diskuteres resultatene

Til sist blir resultatene av undersøkelsene oppsummert og diskutert i kapittel 6. Resultatene blir diskutert i forhold til hverandre og i et teoretisk perspektiv. Deretter presenteres konklusjonen og forslag for videre forskning.

Kapittel 2

Stilometrisk teori og metode

Dette kapittelet vil danne det teoretiske fundamentet for resten av studien. Først vil jeg gi en kort oversikt over stilometri og den historiske utviklingen av fagdisiplinen. Deretter kommer jeg til å greie ut om hvordan stilometriske metoder utføres, med valg av problemstilling, trekk og statistiske metoder. Deretter vil jeg gjøre rede for kritikk av stilometriske metoder og forskning.

2.1 Stilometri - generelt og historisk utvikling

En egenskap som karakteriserer stilometri er flerfaglighet. Litteratur, lingvistikk, statistikk og psykologi er alle fagfelt stilometri kan anvendes innenfor. Når noen skal forsøke å fastslå forfattere av en kjent tekst, med ukjent eller omstridt forfatter(e) kombineres stilometri med litteratur. Valget av trekk for å gjenkjenne forfattere kan være lingvistiske. Eksempelvis kan trekkene være ordbaserte, semantiske og syntaktiske. Statistiske metoder blir brukt til å håndtere frekvensene av trekkene. Om man undersøker om skrivestil endrer seg med alderen, ved demens eller om personlighet innvirket på skrivestilen kommer man innpå de psykologiske eller kognitive fagfeltene.

Stilometri forutsetter at skrivestil kan måles, karakteriseres og gjenkjennes statistisk. Det forutsetter i tillegg at skrivestil, til en viss grad, er underbevisst.

Det finnes en hypotese som heter *the human stylome-hypothesis*. Den går ut på at alle forfattere har et eget *fingeravtrykk* (Van Halteren et al., 2005, s. 65): "(...) authors can be distinguished by measuring specific properties of their writings, their stylome as it were.".

Hypotesen forutsetter at skrivestil er en grad av underbevissthet. Et eksempel er hvis vi forutsetter at en forfatter kan gjenkjennes ut i fra adverbfrekvenser. Dersom noen ikke vil bli gjenkjent kunne de unngått eller brukt flere adverb for å forkle skrivestilen sin. Dette er mulig hvis vi er kjent med hvilke trekk som karakteriserer en skrivestil og vi kan dermed manipulere en tekst til det ugjenkjennelige. Sannsynligvis er enkelte stilmessige trekk vanskeligere å forkle enn andre. Eksempelvis er funksjonsord høyfrekvente og er antatt å være lite sannsynlige at forfattere kan manipulere i en tekst (Nerbonne, 2007). Å endre setningsstruktur i en tekst er sannsynligvis mer krevende enn å endre leksikalske formuleringer, som å bytte "TV" med "fjernsyn".

Det er gjort arbeid med manipulering av tekster for å undersøke om forfatterskap kan forkles

eller imiteres. *Adversarial* stilometri utfordrer antagelsen om hvor underbevisst skrivestil er. *Adversarial* stilometri defineres som bruken av forkledning av skrivestilen, for å endre utfallet av en stilometrisk analyse (Brennan et al., 2012). *Adversarial* stilometri er kommet tilstrekkelig langt til at manuell manipulering av skrivestil gjør det mulig å redusere nøyaktigheten til en statistisk modell til et nivå tilsvarende tilfeldige gjetninger (Brennan et al., 2012). Automatisk forkledning, som maskinoversettelse er ennå utilstrekkelig for å forkle forfatterskap (Brennan et al., 2012). *Adversarial* stilometri presterer tilstrekkelig til at *the human stylome*-hypostesen svekkes. Det er mulig å manipulere skrivestilen til det ugjenkjennelige, i motsetning til et fingeravtrykk. Den praktiske konsekvensen av dette er at stilometri bør brukes med forbehold i juridiske og påtalemessige sammenhenger.

Den andre delen av *the human stylome*-hypotesen er målbarheten til egenskapene i en persons skrivestil. Innen stilometri må egenskapene, her kalt trekk, kunne måles kvantitativt. Trekkene må kunne detekteres og deretter kvantifiseres for å utføre en stilometrisk undersøkelse, i motsetning til kvalitative metoder. Kvalitative metoder er ofte brukt i forensisk lingvistikk (McMenamin, 2002). Forensisk lingvistikk defineres her som den delen av det lingvistiske feltet som omhandler språk i sammenheng innen det kriminaltekniske og det juridiske feltet. Eksempler på hvor forensisk lingvistikk anvendes er ved undersøkelser av trussel- og selvmordsbrev. I forensisk lingvistikk er ikke alltid kvantitative metoder praktiske å bruke alene, men kan komplementere kvalitative metoder (McMenamin, 2002). Hvis et trusselbrev er kortfattet og håndskrevet ville kvalitative metoder trolig være mer informative enn kvantitative metoder. De kvalitative metodene kan for eksempel undersøke håndskrift. Eksempelvis ville en japaner skrevet et 7-tall på en annen måte enn en nordmann. En japaner ville ikke hatt en strek på midten av tallet, men en liten strek nedover på tuppen av den øvre linjen. En nordmann ville mest sannsynlig skrevet 7 med en strek gjennom tallet på midten, som vist i det nedenstående bildet.

Figur 2.1: Bilde av forskjellige skriftlige uttrykk av 7-tallet [1]

Håndskrift er en karakteristikk som er best målbart manuelt. Stilometri i sammenheng med *the human stylome*-hypotesen begrenser trekkene til trekk som er detekterbare og kvantitative.

En problemstilling innen stilometri er om skrivetil er statisk eller om den endrer seg ved alder og/eller erfaring (Daelemans, 2013). Tidligere funn har konkludert at med en økt alder bruker forfattere flere positivt ladede ord og færre negativt ladede ord og færre selvreferanser (Pennebaker and Stone, 2003). Andre har konkludert med at skrivestil endrer seg lite etter man er fylt 30 år (Nguyen

[1] <http://www.dirtycarsmillioncows.com/wp-content/uploads/2015/02/7.jpg>

et al., 2013). Til sist kan kognitiv funksjon ha en innvirkning på skrivestil. En studie fant redusert og vagere vokabular og redusert syntaktisk kompleksitet hos forfattere etter at de utviklet demens (Hirst and Wei Feng, 2012).

Stilometri oppstod med den engelske logikeren Augustus de Morgan (Zheng et al., 2006). Han foreslo at forfattere kunne gjenkjennes ut fra ordlengde i tekstene: "I would have Greek, Latin and English tried, and I should expect to find that one man writing on two different subjects agrees more nearly with himself than two different men writing on the same subject" (De Morgan, 1882, s. 216).

Hypotesen ble testet av Mendenhall (1887) som sammenlignet tekster fra Atkinsons, Dickinson, Thackaray og John Stuart Mill. Han konkluderte med at ordlengde kunne brukes til å skille mellom forfatterne (Mendenhall, 1887).

Mendenhall (1887) kom dermed med et av de første bidragene innefor stilometri vad å undersøke om gjennomsnittlig ordlengde kan skille forfattere fra hverandre. Andre samtidige bidrag, var Mascol (1888) som undersøkte gjennomsnittlige ordlengder og distribusjon av ordlengder og Sherman (1888) som undersøkte setningslengder i engelsk prosa.

Et annet tidlig og viktig bidrag til fagdisiplinen var av (Mosteller and Wallace, 1964), som undersøkte forfatterskap til *The Federalist Papers*, en samling artikler og essayer, skrevet av tre forskjellige forfattere, men hvor det var ukjent hvem som hadde skrevet hvilke av tekstene. I denne studien ble ord og ordfrekvenser analysert for å angi forfatterskap.

Etterhvert som datamaskiner har gjort det mulig å undersøke større tekstmengder og automatisere prosessen er feltet etterhvert blitt dominert av teknikker fra kunstig intelligens, nevrale nettverk og statistisk mønstergjenkjenning (Brennan et al., 2012).

2.2 De tre stilometriske stegene

Stilometriske undersøkelser kan deles inn i tre valg: problem, trekk og statistisk metode (Zheng et al., 2006). Valgene vil her bli brukt til å gi en teoretisk innføring av hva som karakteriserer en stilometrisk analyse

2.2.1 Forfatteranalyse

Forfatteranalyse er en samlebetegnelse på hva man undersøker innenfor stilometri og er det første steget. Forfatteranalyse er valget av målet ved en undersøkelse. Eksempelvis kan forfatteranalyse være en identifisering av en forfatter eller alderen til en forfatter.

Forfatteranalyse kan deles inn i tre forskjellige typer: Forfatteridentifisering, forfatterkarakterisering og likhetsdetektering (Zheng et al., 2006).

Forfatteridentifisering defineres som den forfatteren som er statistisk sannsynlig å ha forfattet en tekst (Zheng et al., 2006). Dette gjøres ved en sammenligning av andre tekster av samme forfatter. Det blir også kalt *authorship attribution*, som jeg har oversatt til forfatterattribuering[2]. Jeg velger å

[2]Mye av faglitteraturen jeg har brukt har vært engelsk og mangler tilsvarende begrep og litteratur på norsk. Jeg har valgt å oversette enkelte begrep til norsk, deriblant stilometri og forfatterattribuering. En ordliste med oversettelser er

bruke begrepet forfatterattribuering, fremfor *author identification*, fordi "identifisering" impliserer en visshet om forfatteridentitet. Metodene angir en *statistisk sannsynlig* forfatter og derfor mener jeg at forfatteridentifisering er et altfor sterkt begrep, som kan virke villedende.

Forfatterkarakterisering defineres som gjenkjenning av karakteristikker til en forfatter, som kjønn, utdanning og kulturell bakgrunn (Zheng et al., 2006). Tidligere har utdannelse blitt foreslått karakterisert ut fra skrivestil (Corney et al., 2002). Kjønn og alder er tidligere blitt karakterisert i en rekke stilometriske studier (Corney et al., 2002; Koppel et al., 2002; Argamon et al., 2003; Goswami et al., 2009; Peersman et al., 2011; Nguyen et al., 2013). Personlighetstyper er også blitt forsøkt detektert via tekst og personlighetstrekkene introvert/ekstrovert kunne identifiseres stilometrisk til en viss grad (Luyckx and Daelemans, 2008b).

Likhetsdetektering defineres som tekster som undersøkes om de er skrevet av samme forfatter eller ikke (Zheng et al., 2006). Forfatteren(e) trenger ikke være kjent fra før for å gjennomføre en likhetsdetektering. Denne kategorien av forfatteranalyse brukes ofte for å oppdage plagiering (Zheng et al., 2006). Ved å detektere likhet mellom to tekster er det mulig å oppdage om tekstene forfattermessig samsvarer med hverandre.

Valget av oppgave påvirker hvilke tekster som skal danne studiens korpus. Hvis vi skal undersøke alder er det en fordel med et datasett med aldersspredning mellom forfatterne. Datasettet bør inneholde tekster fra en rekke forskjellige forfattere for å minimere sjansen for kategorisering ut fra forfatter og ikke egenskapene vi vil identifisere. I et forsøk med forfatterattribuering bør et korpus, som et minimum, inneholde minst én tekst av hver forfatter vi undersøker og en ukjent[3] tekst som skal attribueres til en av forfatterne.

Tekstlengde må tas i betraktning ved valg av tekstsamling. En tekstgrense på 1000 ord er tidligere blitt foreslått for å danne et korpus for å måle lingvistisk variasjon (Biber, 1990, 1993a). Derimot kan trekk som forekommer sjeldent kreve lengre tekster (Biber, 1993a). Trekk kan deles inn i flere typer og beskrives i neste del.

2.2.2 Trekk

Det andre valget i en stilometrisk undersøkelse er valget av trekk, på engelsk kalt *features*. Trekk er markørene som skal kunne diskriminere mellom kategoriene (f. eks. forfattere) våre. Det stilles to krav til trekk som skal anvendes i stilometriske forsøk. Trekkene må være både tilstrekkelig frekvente til å gi statistisk utslag og kunne måles kvantitativt (Oakes, 2014, s. 5).

Trekk kan deles inn forskjellige typer. Jeg har valgt å dele inn trekkene i 4 kategorier, basert på en inndeling av Stamatatos (2009): leksikalske, syntaktiske, applikasjonsspesifikke og semantiske trekk. Forskjellen mellom Stamatatos (2009) og her er at jeg slår sammen kategoriene som er ordbaserte og tegnbaserte til leksikalske trekk. Grunnen til dette er at kategoriene har flere fellestrekk. Kategoriene bruker de samme verktøyene for å hentes ut, ofte krever de den samme preprosesseringen og begge bruker n-grammer.

lagt til i vedlegg G på s. 113

[3] Teksten trenger ikke ha ukjent forfatter i streng forstand. Å vite forfatteren på forhånd gjør det mulig å vite om modellen vår forutsier riktig forfatter eller ikke.

Trekkene som blir nevnt her er ikke en utfyllende liste. Det er blitt foreslått over 1000 forskjellige trekk (Abbasi and Chen, 2008). Trekktyper blir ofte brukt i kombinasjon med hverandre i forfatterattribueringsforsøk (Stamatatos, 2009). Det finnes ikke konsensus om hvilke trekk eller sett med trekk som er de beste stildiskriminerende markørene (Abbasi and Chen, 2008).

Leksikalske trekk

Leksikalske trekk er her definert som trekk basert på ord eller tegn. Eksempler på ordbaserte trekk er ordlengde, setningslengde, ordrikhet, ordfrekvenser, ord n-grammer og skrivefeil (Stamatatos, 2009). Ordbaserte trekk er de første som ble tatt i bruk i stilometrisk analyser (Mendenhall, 1887; Sherman, 1888). Ordlengde og setningslengde er i ettertid blitt brukt til lesbarhetsanalyser (Flesch, 1948). En av lesbarhetsanalysene som bruker de nevnte trekkene er Flesch-Kincaid (Kincaid et al., 1975). Eksempler på tegnbaserte trekk er tegntyper (f. eks. bokstaver og tall), tegn n-grammer og kompresjonsmetoder (Stamatatos, 2009). N-grammer er inndeling av ord og tegn ut i fra et visst antall tegn. For eksempel kan ordet "piraten" bli delt inn sekvensielle 3-grammer: "pir", "ira", "rat", "ate" og "ten". Frekvensene av 3-grammene kan deretter bli brukt i en forfatteranalyse, hvor frekvensen av 3-grammet er med på å indikere forfatter. N-grammer kan inneholde leksikalsk informasjon (ord: "der" og "kan") og kontekstuell informasjon (f. eks. "og_m", som indikerer koordinering i en setning), bruk av punktum og andre tegntyper (f. eks. store bokstaver og komma) (Stamatatos, 2009). I tillegg håndterer n-grammer støy, som f. eks. skrivefeil og kan være nyttige i språk hvor ordinndeling er vanskelig å utføre automatisk (Stamatatos, 2009). Eksempelvis er orddeling i japansk vanskelig fordi språket ikke har mellomrom for å skille ord fra hverandre (Matsuura and Kanada, 2000).

Med kompresjon menes det at tekster blir komprimert til mindre filstørrelser. Kompresjonsmetoder forsøker å redusere redundans av tegn eller ord i en fil statistisk. Graden av kompresjon kan antas å være karakteristisk til teksten og/eller forfatteren. For eksempel hvis en forfatters tekst komprimeres med 40 %. En annen tekst av samme forfatter kan antas å kunne komprimeres i omtrent tilsvarende grad, men en annen forfatter har enn komprimeringsgrad på 60 %. Kompresjonsmetoder er tidligere brukt til å oppdage plagiering og duplikater (Khmelev and Teahan, 2003).

Syntaktiske trekk

Syntaktiske trekk er ansett til å være mer underbevisst og derfor mer pålitelige trekk enn ordbaserte trekk (Stamatatos, 2009). Syntaktiske trekk er antatt å være påvirket av forfatterens underbevisste vaner for setningsstruktur (Stamatatos, 2009) .

Eksempler på syntaktiske trekk er funksjonsord, ordklasser, *chunks*, setnings- og frasestruktur og syntaktiske feil.

Syntaktiske trekk kan deles i to typer, ut i fra hvor mye prosessering som behøves å ekstrahere trekkene ut. De *grunne*, oversatt fra *shallow*, trekkene krever mindre prosessering, som for eksempel ordklassetagging, setnings- og fraseinndeling og tekstinndeling (Stamatatos, 2009). Motsetningen til grunne trekkk finnes *dype* trekk. Dype trekk krever mer kompliserte prosessering, som ved en

mer fullverdig syntaktisk parsing (Stamatatos, 2009). Et eksempel på syntaktisk parsing kan være identifisering av argumentstruktur.

Funksjonsord kan defineres som ord som uttrykker grammatiske eller syntaktiske forhold i en setning. Funksjonsord står i kontrast til innholdsord, som gir informasjon om hva en setning handler om, f.eks. substantiver. Eksempler på funksjonsord på norsk er deteminativer, konjunksjoner, preposisjoner og adverb. Funksjonsord er et trekk blitt brukt med stor suksess (Stamatatos, 2009). Et av de tidligste forsøkene med funksjonsord var Mosteller and Wallace (1964), der kalt *filler words*, brukte ord som *by*,*from* og *to* for å skille forfatterne Madison og Hamilton av *The Federalist Papers* fra hverandre. Det mest effektive ordet for å skille forfatterne fra hverandre viste seg å være *upon* (Mosteller and Wallace, 1964). Fordelen med funksjonsord er at de er høyfrekvente, de er tilstrekkelig innholdsfrie til å ikke variere temamessig og det antas å være usannsynlig at funksjonsord kan kontrolleres bevisst (Koppel et al., 2009).

Frekvenser av omskrivningsregler er blitt introdusert som syntaktiske trekk, i forfatterattribueringsforsøk (Baayen et al., 1996; Gamon, 2004). En omskrivningsregel kan eksempelvis være:

a. VP -> VP NP

b. VP -> VP

Med en omskrivningsregel menes det at de ovenstående reglene til høyre kan omskrives til de venstrestående VP-uttrykkene. Ved å måle frekvensene av de forskjellige omskrivningsreglene kan man bruke dem som trekk. Et eksperimentet av Baayen et al. (1996) konkluderte med at omskrivningsregler kunne være mer pålitelige enn ordbaserte trekk til forfatterattribueringsformål. I Gamon (2004) presterte omskrivningsregler lavere enn funksjonsord og trigrammer av ordklasser til å forutsi forfatter. Trekkene forutsetter et syntaktisk annotert korpus.

Med utvikling av pålitelige syntaktiske verktøy er automatisk ekstrahering av flere syntaktiske trekk blitt mulig. Syntaktiske trekke har vist seg å forbedre resultatene av forfatteranalyser sammen med ordbaserte trekk (Stamatatos et al., 2001, 2000; Baayen et al., 1996; Gamon, 2004; Van Halteren, 2004; Chaski, 2005; Uzuner et al., 2005; Hirst and Feiguina, 2007; Koppel et al., 2009)

Generelt er syntaktiske trekk språkavhengige og krever mer datamaskinelle prosessering av tekster enn leksikalske trekk. Fordelen er at trekkene ofte er "innholdsfrie" , som vil si at de ikke er påvirket av tema i en tekst men av forfatternes formuleringsvaner (Li et al., 2006).

Semantiske trekk

Semantiske trekk defineres her som trekk som viser til innhold i tekst eller viser til semantiske relasjoner. Eksempler på semantiske trekk er synonymer, tematiske roller og argumentstruktur. (Stamatatos, 2009). Forholdet mellom to synonymer og forholdet i bruken av dem kan karakterisere en forfatter. For eksempel kan en forfatter bruke ordet "TV" i stedet for "fjernsyn"[4]. Med tematiske roller menes det at frekvensene av de forskjellige tematiske rollene kan være hvert sitt trekk, eksempelvis agens og patiens.

[4]Gitt at ordene har en lik og fakultativ distribusjon, for å få et binært forhold mellom dem.

Med argumentstruktur menes antall argument et predikat krever. Eksempelvis kunne en forfatter brukt forskjellige predikat, som i gjennomsnitt tok 2.3 argument i tekstene hans. En annen forfatter kunne hatt en predikatfrekvens på 1.7 gjennomsnittlige argument. Frekvensene kan dermed brukes til å skille mellom forfatterne.

Semantiske trekk, nærmere bestemt funksjonstrekk, ble ekstrahert i en forfatteranalysel av Argamon et al. (2007). Funksjonstrekk kan defineres som ord eller fraser valgt ut i fra en semantisk-funksjonell analyse (Argamon et al., 2007). Trekkene ble delt inn i tre typer: *cohesion*, *appraisal* og *assessment* (Argamon et al., 2007). For å få en bedre forståelse for typene inkluderte for eksempel *cohesion* en undergruppe med konjunksjoner. Konjunksjonene ble delt inn i tre undergrupper: *Elaboration* (f. eks. *that is, rather*), *Extention* (f. eks. *and, or, but, yet*) og *Enhancement* (f. eks. *then, next, similarily*). Eksperimentet har vist at funksjonstrekk kan bidra i oppgaver med tekstklassifisering (Argamon et al., 2007).

Et forbehold med semantiske trekk er om de er innholdsavhengige eller ikke, eller i hvor stor grad de er innholdsavhengige. For eksempel er synonymer ofte innholdsord. Unntaket er funksjonsord som er syntaktiske og kan være innholdsfrie. Semantiske trekk kan kreve en større grad av prosessering for å hentes ut, i likhet med syntaktiske trekk.

Applikasjonsspesifikke trekk

Applikasjonsspesifikke trekk defineres her som hvordan forfattere organiserer en tekst og tekstavhengige trekk. Eksempler på applikasjonsspesifikke typer trekk er strukturelle trekk, innholdsspesifikke trekk og språkspesifikke trekk (Stamatatos, 2009).

Strukturelle trekk omhandler oppsettet av en tekst, blant annet avsnittslengde, kursivering, hilsener, avskjeder og signaturer (De Vel et al., 2001; Li et al., 2006; Teng et al., 2004; Zheng et al., 2006).

Innholdsspesifikke trekk defineres som frekvenser av nøkkelord. Når tekster er kontrollert for tema og sjanger kan ord som er høyfrekvente i tekstene velges for å karakterisere forfattere, kalt nøkkelord (Stamatatos, 2009). Et eksempel på nøkkelord var en kriminell selger som brukte betegnelsen, *obo* (=*or best offer*) ved salg av en piratkopiert programvare og ble gjenkjent for forkortelsen (Zheng et al., 2006). Trekkene kan brukes i et komplimentært forhold til innholdsfrie trekk for å forbedre nøyaktigheten i enkelte forsøk (Li et al., 2006).

Eksempel på språkspesifikke trekk er diglossia. Fenomenet ble undersøkt i moderne gresk, hvor verbendelsene kan deles inn mellom uformel og formel gresk og ved hjelp av de forsøke å skille forfattere fra hverandre (Tambouratzis et al., 2004). Alene kunne ikke diglossia for gresk skille mellom forfattere (Tambouratzis et al., 2004).

Språkspesifikke trekk kan være relevant for norsk, som har en formel skriftmålsinndeling: bokmål og nynorsk. I tillegg finnes subnormer, som konservativ og moderat bokmål. Valget mellom bokmål og nynorsk kan sies å være i høy grad et bevisst valg, men trolig er subnormer i mindre grad bevisst valgt. En undersøkelse av normklynger indikerte et implikasjonshierarki mellom a-endelser og en-endelser. For eksempel vil noen som bruker order “avisen” mest sannsynlig også bruke en-

endelse ordene "tiden" og 'høyresiden" (Dyvik, 2012, s. 208). Databaser og skriverettingsprogram som inneholder informasjon om subnormer, eksempelvis SCARRIE (De Smedt and Rosén, 1999) for norske subnormer, kunne vært et aktuelt verktøy for en subnormrettet stilometrisk undersøkelse.

2.2.3 Statistiske metoder

Det siste valget i en stilometrisk undersøkelse er valget av statistisk metode. Statistiske metoder kan deles inn i ikke-overvåkede og overvåkede metoder. Ikke-overvåkede metoder defineres her som statistiske metoder som forsøker å klassifisere instanser ut i fra likheten til hverandre. Overvåkede metoder defineres her som statistiske metoder som deler instanser inn etter forhåndsbestemte kategorier.

Inndelingen reflekteres i Sebastiani (2005), som definerer forfatteranalyser som et tekstklassifiseringsproblem. Tekstklassifisering blir delt inn i to forskjellige metoder for å klassifisere tekster på: *text clustering* og tekstkategorisering. *Text clustering* forsøker å oppdage klassestruktur i en tekstsamling (Sebastiani, 2005). Tekstkategorisering forsøker å klassifisere innen ett gitt skjema (Sebastiani, 2005).

En annen inndeling, av Zheng et al. (2006), delte statistiske metodene inn i to undergrupper: Statistisk analyse og maskinlæring (Zheng et al., 2006). Inndelingen kan være lite presis, fordi kategoriene overlapper. For eksempel kan evklidisk distanse anvendes med både maskinklæringsalgoritmen *k-Nearest Neighbour* (Calix et al., 2008) og statistiske multivariate analyser (Dabagh, 2007).

Denne inndelingen illustrerer et skillet i stilometrisk forskning som skjedde med inntoget av forskere med datamaskinell- og makinlæringsbakgrunn. Inndelingen overlapper delvis med ikke-overvåkede/overvåkede metoder, hvor maskinlæringsmetodene ofte er overvåkede metoder (f.eks. SVM, k-NN, *Naive Bayes* og *decision trees*) og de statistiske metodene (f. eks. *Cluster Analysis*, *Principal Component Analysis* og *Factor Analysis*) ofte er ikke-overvåkede metoder. Jeg kommer til å bruke inndelingen med statistisk analyse/maskinlæring videre for å gruppere de forskjellige statistiske metodene.

Avslutningsvis kan det sies at det mest sentrale punktet er ikke inndeling av de statistiske metodene, men hvordan man kommer frem til den beste måten for å klassifisere tekstene ut ifra problemstillingen man velger.

Multivariate statistiske analyser

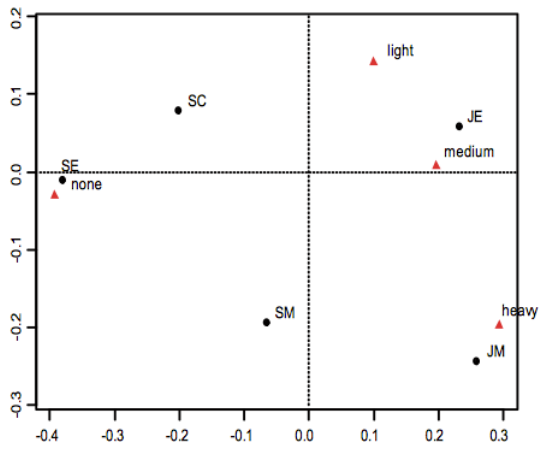
I de tidligste forsøkene var de statistiske metodene begrenset til å håndtere kun ett eller to trekk (Zheng et al., 2006). For eksempel brukte Mendenhall (1887) kun trekket ordlengde for å undersøke forskjeller mellom forfattere. Forskningen har utviklet seg fra å bruke univariate modeller som bruker én variabel til multivariate modeller som håndterer flere variabler.

Holmes (1994) nevner fire typer multivariate statistiske metoder som er brukt i de senere årene: *Factor Analysis*, *Discriminant Analysis*. *Cluster analysis* og *Principal Component Analysis*. Andre statistiske metoder som også er brukt i stilometriske undersøkelser er *Multidimensional Scaling*

(López-Escobedo et al., 2013) og *Correspondence Analysis* (Tabata, 2007). Jeg vil gi en kort oversikt over hver av de nevnte statistiske metodene og stilometrisk arbeid med metodene.

- *Correspondence Analysis* (CA)

  CA er en statistiske visualiseringsmetode for å illustrere en sammenheng mellom rader og kolonner i en tabell (Young and Bann, 1996). I stilometriske undersøkelse vil en CA ha en tabell med frekvensene av kategoriene på en side og trekkene på en annen side. Deretter blir distansen regnet ut mellom kategoriene og trekkene. Resultat kan visualiseres i en todimensjonal figur, som figuren 2.2 til venstre. I figur 2.2 indikerer trekantene hvor mye ansatte røyker og prikkene er titler til ansattkategorier (f.eks SE = *Secretaries*). Linjene kan tolkes som at det er et skille mellom *none* og *medium* røyking og *light* og *heavy* røyking. Eksempelvis er sekretærer (SC) plassert i et kvadrat er på *none*-siden og *light*-siden av røyking. Dette korrensponderer med datasettet til høyre i 2.2, hvor flest sekretærer er ikke-røykere, i forhold til medium-aksen og det er flere tilhører *light*-kategorien fremfor *heavy*-kategorien.



(a) CA-eksempel med ansatte og røykere

```
> smoke
   none light medium heavy
SM    4     2      3     2
JM    4     3      7     4
SE   25    10     12     4
JE   18    24     33    13
SC   10     6      7     2
```

(b) Datasett med ansatte og røykere

Figur 2.2: *Correspondence Analysis*eksempel [5]

I Tabata (2007) ble en studie av superlativene i verkene til Dickens og Smollet undersøkt med CA. Superlativene ble kvantitativt sammenlignet mellom forfatterne for å undersøke korrelasjoner mellom tekst og superlativfrekvenser. Resultatene viste at CA kunne skille mellom forfatternes frekvensene av superlativer og at begge forfatterne generelt hadde lavere frekvenser av superlativer i tidligere verk.

Jamfør Linmans (1998) og Mealand (1999) for flere studier med CA og stilometri.

- *Factor Analysis* (FA)

  FA bygger på parvise korrelasjoner mellom variablene for å identifisere et mindre sett av underliggende strukturer (Biber, 1993a). FA kan deles inn i *exploratory* og *confirmatory*. I *exploratory*

[5] <http://statmath.wu.ac.at/courses/CAandRelMeth/caipB.pdf>

faktoranalyse forsøker modellen å finne en underliggende struktur. *Confirmatory* faktoranalyse blir brukt til å teste en hypotese om underliggende strukturer eller dimensjoner (Torres-Reyna, 2010).

Biber (1993a) brukte FA til å undersøke kollokasjonene, *certain* og *right*. FA viste seg å være nyttig for å oppdage underliggende mønster i kollokasjoner (Biber, 1993a).

- *Discriminant Analysis* (DA)

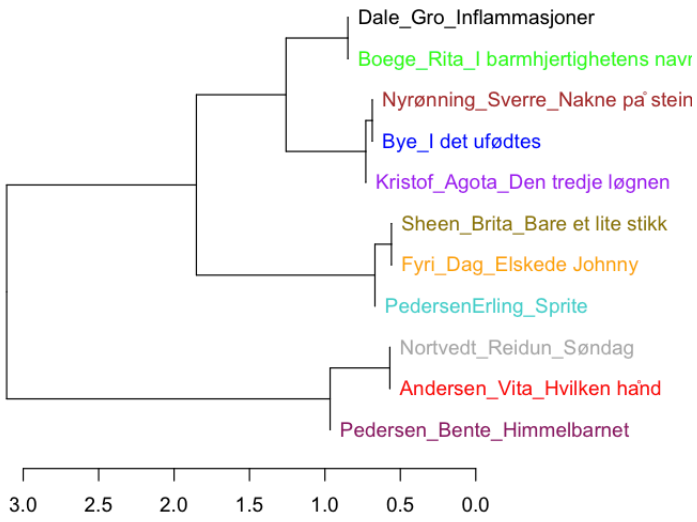
  DA forsøker å dele inn instanser (f. eks. tektser) ut i fra naturlige grupper (f. eks. forfattere). Ut fra frekvensene som er valgt til å representere en tekst forsøker DA deretter å plassere teksten i en av gruppene, i dette tilfellet den forfatteren som frekvensen ligger nærmest.

  DA kan brukes når tekster på forhånd er delt inn i grupper av forfattere (Holmes, 1994). DA klassifiserer instansene i grupper ut i fra hvor lik instansene er de tidligere definerte gruppene. DA er en 2-stegsprosess. Først testes signifikansen på et sett med diskriminerende funksjoner og deretter klassifiseres instansene (Poulsen and French, 2008).

  I et forsøk av Baayen et al. (2002) ble DA foretrukket fremfor *Principal Component Analysis* når det gjaldt tekster av en ukjent forfatter som ble testet på et treningssett med tekster av kjente forfattere.

- *Cluster Analysis* (CLA)

  CLA forsøker å identifisere grupper og plassere objekter inn i gruppene ut i fra likhet mellom objektene. Med CLA kan man formere grupper med relaterte variabler, på samme måte som i *factor analysis*(Norusis, 2008). CLA egner seg i forsøk der man på forhånd ikke vet gruppetilhørighet, eller hvis man vil bekrefte gruppetilhørighet. Figuren 2.3 gir et eksempel på CLA visualisering og grupperinger i CLA. I figuren er 11 forfattere og tekster kategorisert ut fra nærhet ut fra de 100 mest frekvent 6-grammene. Den nedre linjen som går fra 3.0-0.0 viser til avstand mellom klyngene, f. eks. er Dale (øverst) nærmere Bye, enn Pederson sett ut fra hvor langt man må gå bakover i figuren for å komme til riktig klynge.



Figur 2.3: CLA-eksempel i et dendrogram, laget av R-pakken Stylo

CLA er tidligere anvendt i stilometriske undersøkelser. Hoover (2001) brukte frekvenser av høyfrekvente ord av kjente forfattere for å undersøke nøyaktigheten til CLA. CLA hadde resultat med en nøyaktighet på mindre enn 90 %. Hoover (2001) mente de dårlige resultatene til CLA indikerte generelle problemer og ikke forsøksspesifikke problemer ved bruk av CLA til forfatterattribuering.

- *Principal Component Analysis* (PCA)

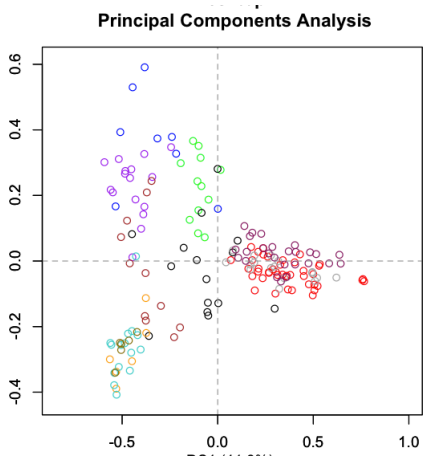
  PCA forsøker å identifisere mønstre i et datasett. Mønsteret blir uttrykt ved å vise til likhetene og forskjellene i datasettet. PCA kan finne mønster i datasett med høye dimensjoner og redusere dimensjonene ved kompresjon (Smith, 2002). Med andre ord reduserer PCA dimensjonaliteten i et problem. Figur 2.4 likhetene mellom 11 forfattere sett ut fra de 100 mest frekvente 6-grammene. Fargene viser forskjellige tekster og prikkene forskjellige deler av tekstene delt inn i deler av 200 ord hver..

  Figur 2.4 er et eksempel på visualisering i PCA.

  PCA er blitt brukt i en rekke stilometriske undersøkelser (jfr. Burrows, 1987; Baayen et al., 2002, 1996; Burrows, 1992)

  En senere studie med høyfrekvente funksjonsord har indikert at PCA kan tidligere ha vært overvurdert i forsøk:

  > "..our experimental texts fails to uncover authorial structure suggests that the authors studied in literary stylometry, for which principal components analysis is reported to lead to insightful clustering (...) These are authors who must have developed their own writing style far beyond the more rudimentary differences in style that we could only observe for our participants by using far more powerful analytical tools than simple principal components analysis..." (Baayen et al., 2002, s. 37)



Figur 2.4: PCA-eksempel laget i R-pakken *Stylo*

- *Multidimensional Scaling* (MDS)

  MDS forsøker å gruppere etter likheter og forskjeller ved å plassere instansene i riktig lengde fra hverandre i kategorier som på forhånd ikke er gitt. "Multi" referer til at distansene kan plasseres i flere dimensjoner enn to. MDS kan sees som et alternativ til FA (StatSoft, 2013). FA analyserer likheter mellom instanser og uttrykker dette i en korrelasjonsmatrise. MDS kan også danne korrelasjonsmatriser som analyserer i likhets- eller ulikhetsmatriser.

  MDS er brukt i forsøk på to spanske korpus med korte og lange tekster. Målet var å undersøke egenskapene til trekkene i et forfatterattribueringsforsøk med et korpus som varierte i tekstlengde og sjanger (López-Escobedo et al., 2013). Resultatene for korte tekster var mindre nøyaktige enn for lengre tekster og dannet ikke like klare klynger. Dette kan skyldes mangel på representativitet av trekkene eller at de lange tekstene var skrevet av profesjonelle forfattere (López-Escobedo et al., 2013). En litterær forfatterstil har vist seg å påvirke resultatene med PCA (Baayen et al., 2002), som er en problemstilling som er potensielt overførbar til andre statistiske metoder.

Maskinlæringsmetoder

I de senere årene har maskinlæring blitt svært populært å anvende innenfor stilometri. I forfatterattribueringsproblem forsøker maskinlæring å forutsi forfatteren til en tekst, basert på et sett med trekk (Zheng et al., 2006). Med klassifiseringmetoder menes det at klassifikatoren lærer seg forskjellige klasser og klassifiserer instanser ut ifra den mest sannsynlige klassen.

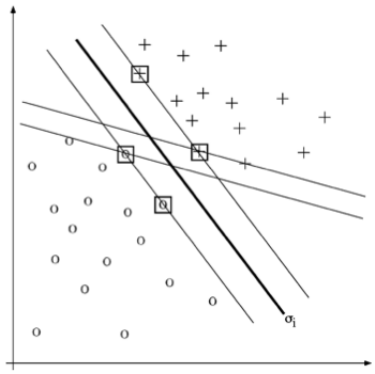
Den enkleste strategien, som ikke lærer fra trekkene, men kun tar i betraktning fordelingen i klasser, velger alltid den mest frekvente klassen. For eksempel har et korpus 10 tekster, hvor 5 er skrevet av Arne, 2 er skrevet av Lise og 3 er skrevet av Bodil. Modellen er programmert til å velge den mest frekvente klassen, hvor klassen i dette tilfellet er en av forfatterne, og den mest frekvente klassen er Arne. Modellen vil deretter klassifisere 50 % riktig, siden Arne har skrevet 5 av 10 tekster og dermed er den mest frekvente klassen. Dette er en regelbasert klassifikator og ikke en maskinlæringsalgoritme. Denne regelbaserte klassifikatoren kan brukes til å etablere en *baseline* i et forsøk, for å kunne evaluere resultatene i forhold til andre klassifikatorer.

(Sebastiani, 2005). Nøyaktigheten til klassifikatorene til maskinlæringsmetodene er etter hvert blitt svært gode, og forbigår regelbaserte modeller (Sebastiani, 2005). Fordelen med et maskinlæringsparadigme som lærer av tidligere eksempler (induktiv læring) er at det er en høyere automatisering av prosessen enn regelbaserte modeller

Det finnes en rekke typer maskinlæringsalgoritmer, de 5 typene jeg nevner videre er ikke en endelig oversikt, men er populære og ofte brukte modeller innen forsøk med maskinlæring og stilometri.

- *Support Vector Machines* (SVM)

  SVM er en relativt ny *supervised learning*-teknikk (Vapnik, 2000). SVM er basert på *marginer*. På hver sin side av et hyperplan deles to dataklasser fra hverandre. Man maksimerer marginen og dermed forsøke å danne størst mulig distanse mellom de delte sidene til hyperplanet og instansene på hver side av dem. Figur 2.5 er et eksempel på et hyperplan, som forsøker å skille kategoriene, kryss og sirkler, fra hverandre. Den tykkeste streken er punktet som er lengst fra vektorene (de firkantede boksene) og skiller hyperplanet med maksimal margin ved hjelp av vektorene.



Figur 2.5: SVM eksempel Sebastiani (2005, s. 30)

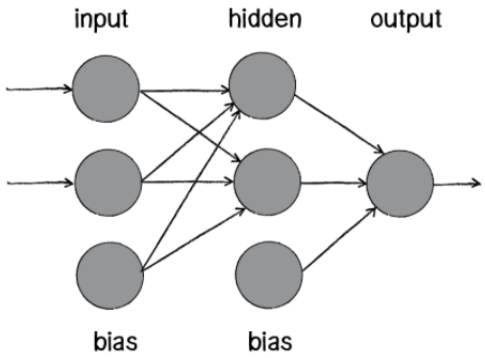
SVM egner seg til oppgaver med mange trekk i forhold til treningsinstanser (Kotsiantis, 2007). Det er fordi SVM kan håndtere høy dimensjonalitet (Stamatatos, 2009). SVM krever en stor mengde treningsdata for å oppnå best mulig klassifiseringsrate (Kotsiantis, 2007).

I Hirst and Feiguina (2007) ble SVM brukt til å undersøke frekvensene til bigram av syntaktiske annotasjoner ved en delvis parsing av tekstene. I denne undersøkelsen var SVM et bedre valg enn PCA og hadde høytere riktig klassifiseringsrate.

- *Nevrale nettverk*

  Nevrale nettverk er basert på idéen om *perceptron* av Rosenblatt (1961). Nevrale nettverk forsøker å håndtere instanser som ikke er lineært separerbare (Rumelhart et al., 1985). Et nettverk kan bestå av *input*-noder, skjulte noder og *output*-noder. Helt enkelt kan det sies at nevrale nettverk lærer ved å ta et treningsett, kjøre settet gjennom nettverket flere ganger, helt frem til algoritmen finner riktig justering av vekten som produserer best output for treningssettet (Kotsiantis, 2007). Nevrale nettverk blir bedre jo lengre det kjører (Kotsiantis, 2007).

  Figuren 2.6 er en illustrasjon av et nevralt nettverk. Det har *input*-noder, skulte noder og *output*-noder. Linjene mellom nodene er vekter. Vektene er styrken mellom nodene. Hvis modellen produserer korrekt *ouput* er det ikke nødvendig å justere på vektingen mellom nodene, men hvis det gir utilstrekkelig *output* kan vektingen justeres. Ved å justere vektingen lærer modellen.



Figur 2.6: Eksempel på et nevralt nettverk.[6]

Dersom man ser på forfatterattribuering som et mønstergjenkjenningsproblem (her synonymt med *text clustering*) har nevrale nettverk evnen til å gjenkjenne underliggende mønstre.

I tidlig stilometrisk forskning med nevrale nettverk ble tekster av Shakespeare og Fletcher undersøkt (Matthews and Merriam, 1993; Merriam and Matthews, 1994). Tekstene var arbeid begge forfatterne var assosiert med. Et av trekkene som ble brukt var frekvensene av ordratio for eksempel: *did/(did+do)*. Et nevralt nettverket forsøkte deretter å gjenkjenne Shakespeare og Fletcher sine arbeid. Ut fra ordratio ble nettverket bedt om å attribuere forfatterskap til flere tekster begge var assosisert ved. Konklusjonen var at enkelte tekster, *Two Noble Kinsmen, Double Falsehood*

[6] <http://natureofcode.com/book/chapter-10-neural-networks/>

og *London Prodigal* var sammarbeid mellom Shakespeare og Fletcher (Matthews and Merriam, 1993).

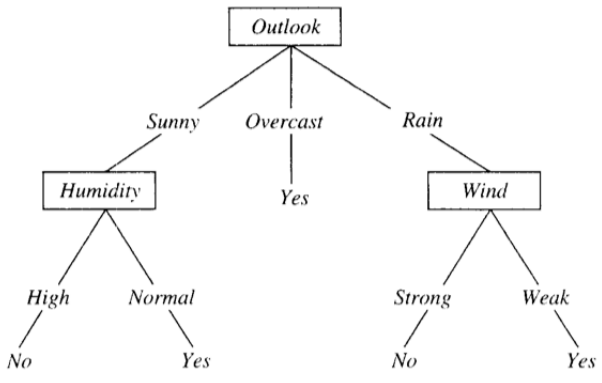
Radial Basis Function er en alternativ type nevrale nettverk, som ved å identifisere sfærer i forskjellige deler av et område kan skille kategorier fra hverandre (Holmes, 1998). I et senere forsøk ble sammarbeidene mellom Shakespeare og Fletcher undersøkt igjen. Lowe and Matthews (1995) sine resultat var forskjellige fra det forrige nevrale forsøket (Matthews and Merriam, 1993). I *Two Noble Kinsman* fikk deler av stykket ikke attribuert forfatter på grunn av for stor usikkerhet om forfatterskap (Lowe and Matthews, 1995).

Generelt presterer nevrale nettverk bedre med kontinuerlige trekk og høyere dimensjoner, enn med diskrete trekk og kategoriske trekk (Kotsiantis, 2007). Sammenlignet med *decision trees* kan nevrale nettverk prestere på likt nivå som *decision trees*, men sjeldent bedre (Eklund and Hoang, 2002; Lim et al., 2000).

- *Decision trees*

  Et *decision tree* har bestemmelsesnoder, som undersøker verdien av et trekk og blad-noder som angir kategori. I figur 2.7 forsøker treet å forutsi om lørdagsmorgener, ut fra været, er egnet til å spille tennis. Man starter ved toppen av noden, som eksempelvis i figur 2.7 er toppnoden *Outlook*, som undersøker verdien til en instans og bestemmer hvilken gren instansen skal klassifiseres til. Dette fortsetter frem til man kommer til en *output*-node, som eksempelvis i figur 2.7 er *Yes* eller *No*.

  Generelt pleier logikkbaserte systemer, som *decision trees* å gjøre det bedre med diskrete eller kategoriske trekk (Kotsiantis, 2007).



Figur 2.7: Eksempel av *decision tree*.[7]

I Dumais et al. (1998) sitt forsøk med forskjellige maskinlæringsalgoritmer konkluderte med at *decision tree* presterte godt i en tekstkategoriseringsoppgave. Trærne presterte lavere enn SVM, og høyere *Naive Bayes* i en tekstkategoriseringsoppgave. De gode resultatene begrunner av at læringsmetoden kan håndtere fleksible og dynamisk informasjon (Dumais et al., 1998).

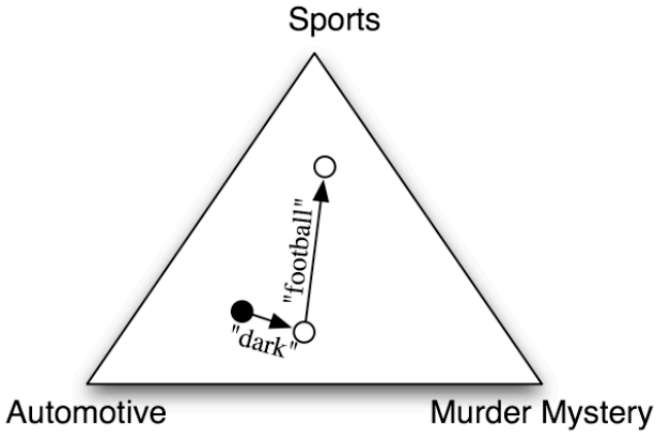
[7] <https://www.cs.princeton.edu/courses/archive/spring07/cos424/papers/mitchell-dectrees.pdf/>

Svakheter med *decision trees* er at de presterer best med få, men svært relevante attributter fremfor mange, komplekse attributter (Rokach and Maimon, 2005). I tillegg er trærne svært sensitive for treningssettet og kan reagere negativt på uviktige attributter og støy i datasettet (Rokach and Maimon, 2005).

- *Naive Bayes* (NB)

  Naive Bayes har en todelt klassifiseringsprosess. Først finner den en tidlig sannsynlighet for kategorisering av en instans ved å regne ut sannsynligheten til hver kategori. Deretter regner NB ut sannsynlig kategorisering av instansen. Sammen med den tidlige sannsynlighetsutregningen, klassifiserer den instansen. NB har en underliggende probabilitetsmodell som forutsetter at trekkene er uavhengige av hverandre og klassifiserer instansene ut fra denne *naive* antagelsen (Bird et al., 2009).

  I figur 2.8 skal tekster klassifiseres etter sjanger ut fra ord. Eksempelvis skal ordet *dark*, klassifiseres. Fra før er flest tekster kategorisert som *Automotive* og har dermed høyest sannsynlig tidlige klassifisering. Det viser seg at ordet er en mindre indikator for kategorien *Murder Mystery*, som forekommer oftere i den kategorien enn i de andre kategoriene og derfor plasseres nærmere. Det samme gjelder ordet *football*, som er en sterk indikator på kategorien *Sports* og nærmer seg den kategorien.



Figur 2.8: Eksempel av NB klassifisering (Bird et al., 2009).

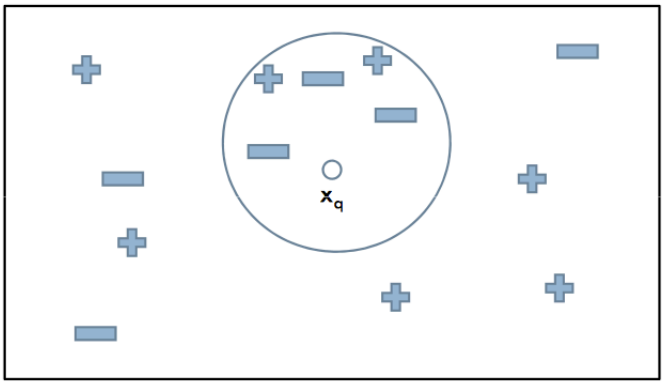
  For numeriske trekk er det vanlig å omgjøre dem til diskrete verdier i en preprosesseringsfase (Yang and Webb, 2003), som også kan bestå av normaldistribusjon av probabilitetskalkuleringen (Bouckaert, 2003). I tillegg trenger NB en relativ liten rate treningsdata for å oppnå optimal klassifiseringsrate i forhold til SVM og nevrale nettverk (Kotsiantis, 2007).

  NB har anvendt i et tidligere i et forsøk med syntaktiske og leksikalske trekk (Koppel et al., 2002). NB fikk lavere klassifiseringsrate enn en annen modell fordi NB ikke tar hensyn til avhengigheter mellom trekk (Koppel et al., 2002).

- *k-Nearest Neighbour* (k-NN)

k-NN baseres på antagelsen at instansene i et datasett kan eksistere i nærhet til andre instanser med like egenskaper. k-NN er en minnebasert læringsalgoritme og kan karakteriseres av *lazy-learning*. *Lazy-learning* utsetter generaliseringsprosessen helt frem til klassifiseringen er ferdig utført. k-NN krever mindre beregningstid i treningsfasen i forhold til *eager-learning algoritmer* (f. eks. nevrale nettverk og NB), men krever mer beregningstid i klassifiseringsfasen (Kotsiantis, 2007). k-NN krever stor lagringskapasitet og er sensitiv når det kommer til valget av likhetsfunksjon, som er brukt til å sammenligne instansene (Kotsiantis, 2007)..

Figur 2.9 illustrerer hvordan k-NN klassifiserer. Modellen har her k=5 og en klassifisering vurderer derfor 5 av de nærmeste instansene av X. I dette tilfellet er det tre minustegn og 2 plusstegn i nærheten. X blir dermed klassifisert som et minustegn.



Figur 2.9: Eksempel av k-NN algoritme [8].

Eksempelvis er k-NN brukt tidligere sammen med en kombinasjon av leksikalske og syntaktiske trekk(Luyckx and Daelemans, 2008b). Modellen kunne forutsi nesten 50 % riktige forfattere, ut av 145 stykker. Trekkene ble modellert i TiMBL, et maskinlæringsprogram for minnebasert læring (Daelemans et al., 2003).

Maskinlæringsprogram

Av dataprogram egnet til stilometriske undersøkelser med maskinlæring er blant annet *TiMBL: Tilburg Memory-Based Learner* (Daelemans et al., 2003), tidligere nevnt i delen om k-NN. Den anvender algoritmer for minnebasert læring.

Programmet er blant annet brukt i stilometriske undersøkeler av Luyckx and Daelemans (2008b) for å forutsi personlighetstrekk ut fra tekster, til undersøkelser av forfattere med begrenset data (Luyckx and Daelemans, 2008a) og til STYLENE, en nettside hvor man skal kunne utføre stilometriske undersøkelser på nederlandsk (Daelemans and Hoste, 2013).

Fordelen med dette programmet er at det håndterer store datamengder og er laget for ikke-diskrete data. Ulempen er at det kreves preprosessering av tekstene i forkant, er begrenset til minnebasert læringsalgoritmer og har ikke mulighet for grafisk fremstillinger av resultatene.

[8] `http://www.csee.umbc.edu/~tinoosh/cmpe650/slides/`

WEKA er et annet program for maskinlæring i Java (Hall et al., 2009). Programmet kan bruke læringsalgoritmer av typene SVM, logikkbaserte algoritmer, regelbaserte algorimer, nevrale nettverk og instansbasert læring. I tillegg er det mulig å velge bort trekk som ikke skal brukes i preprosesseringen og visualisere resultatene.

Programmet er blant annet brukt til *adversial* stilometri i Brennan et al. (2012), i Brocardo et al. (2014) for kontinuerlige autentisering av forfattere og ved forsøk med identifisering av oversettere (El-Fiqi et al., 2011).

Fordelene med WEKA er valgfriheten av algoritmer, mulighetene til å skille trekkene fra hverandre i preprosesseringen og visualiseringene av resultatene. Bakdelen med WEKA er at programmet krever preprosessering av tekstene, blant annet til i riktig format og at WEKA håndterer relativt små datasett i forhold til TiBML, dersom man ønsker grafisk fremstillingen av programmet. Dersom datasettet er stort kan det håndteres i terminalen av WEKA.

2.3 Stilometriskbaserte programvarer

Det finnes etter hvert en rekke programmer som anvender stilometri, enten til stilometrisk forskning eller til andre, allmenne formål.

Det finnes andre program som anvender stilometriske metoder, men de er ofte rettet mot stilometrisk forskning, deriblant JGAAP og *Stylo* (Eder et al., 2013). Disse programmene kan også ha praktiske anvendelser, men er primært laget til forskning. JGAAP ble blant annet brukt i undersøke om *The Cuckoo's Calling* ble skrevet av Rowling, ved å sammenligne boken med en bok av Rowling og bøker til tre andre forfattere (Juola, 2013).

Et program rettet mot et mer allmennt formål er Anonymouth.

Anonymouth

Anonymouth er et program som brukes til å anonymisere skrivestiler, utviklet av Drexel University[9]. Målet er et verktøy for de som ønsker å skjule identiteten sin. Et praktisk eksempel av Anonymouth, er journalister i land uten fri presse.

Anonymouth anvender maskinlæring og metoder innenfor språkprosessering til å anonymisere forfatteren (McDonald et al., 2013). For å skjule forfatteren i en tekst anvender programmet tre forskjellige dokumenter: Et som skal anonymiseres, en eksempel tekst av forfatteren og en siste tekst forfattet av andre forfattere. Programmet fjerner karakteristiske ord ved å oversette til andre ord og gir en liste over forslag til endringer av setninger.

Videre utvikling av Anonymouth blir å automatisere programmet i større grad enn nå, gjøre det mer robust til å håndtere forskjellige situasjoner og øke brukervennligheten til programmet (McDonald et al., 2013).

[9] <https://www.cs.drexel.edu/~pv42/thebiz/>

2.4 Et kritisk blikk på stilometri

Tidligere i kapittelet ble problemstillinger rundt idéen om stilometri diskutert. Jeg vil nå gi en kort oversikt over aktuell kritikk rettet mot stilometri. Mye dette er tidligere kritikk av Rudman (1998, 2010, 2012). Jeg henter inn deler av kritikken relevant for denne studien, men anbefaler en gjennomgang for de virkelig interesserte.

Generelt kan det sies at stilometri preget av en mangel på konsensus. Det er uenighet om resultatene av undersøkelser er definitive, om korrekt metodebruk og statistiske metoder (Rudman, 1998).

Et tidvis problem er at av problemene ved stilometrisk forskning at forskere mangler ekspertise innen deler av det de forsker innen stilometriske analyser, i følge Rudman (1998). Med dette menes det at dersom man inn i fagdisiplinen for å drive forskning må man ha nødvendig ekspertise til det man vil undersøke (Rudman, 1998). Eksempelvis hvis man undersøker 1800-talls litterære verk bør man ha kunnskap om litteratur fra denne perioden og bare ha kunnskap om statistikk. I tillegg finnse det dilettanter som, uten å sette seg inn i disiplinen, gjør en undersøkelse og deretter går videre til noe annet (Rudman, 1998).

Det er skrevet mange artikler, bøker og kapitler om stilometri, men til tross for dette finnes det ikke konsensus om en grunnleggende bibliografi av disiplinen (Rudman, 1998). Dette kan føre til mindre optimale valg av metoder og forvirring rundt konsept og begrep (Todorov, 1971). Eksempelvis ble det tidligere nevnt at forfatteridentifisering og forfatterattribuering kunne brukes synonymt, og at det finnes overlapp mellom begrepene til de statistiske metodene diskutert. Dette kan skyldes at stilometri, som disiplin, favner om flere fagfelt. Det er problematisk fordi det fører til lite koherens i forskningen og at nyere forskning potensielt ikke tar i bruk relevant, tidligere forskning (Rudman, 1998).

Et annet problem er bruk av uegnede statistiske metoder (Rudman, 1998). Med andre ord hevder Rudman (1998) at enkelte bruker statistikk for å bevise noe statistikken ikke er egnet til. Et konkret eksempel på hvordan uegnede metoder kan få alvorlige konsekvenser er CUSUM-kontroversen, som Holmes and Tweedie (1995) gir en grundig oversikt. På tidlig 90-tallet ble en statistisk teknikk, basert på tabeller med kummulative summer av trekk, eksempelvis brukt på ordklasser for å attribuere forfatter (Holmes and Tweedie, 1995). Metoden ble brukt til rettslige og juridiske formål (Holmes and Tweedie, 1995). I England ble den blant annet brukt av forsvarere til å svekke troverdigheten til tilståelsene av klientene deres(Holmes and Tweedie, 1995). Ved å bruke CUSUM-tabellen kunne forsvarerne "påvise" at deler av tilståelsene var påvirket av avhørerne. I ettertid ble metoden kraftig kritisert å mangle vitenskapelig belegg til forfatterattribuering, være subjektiv og være avhengig av kunnskapene til den som tolker skjemaet (Holmes and Tweedie, 1995). Mye av denne kritikken har stammet fra en mangel på objektiv evaluering av metoder (Stamatatos, 2009).

I de senere årene har flere statistiske metoder blitt introdusert til disiplinen, blant annet overvåkede og ikke-overvåkede maskinlæringsmetoder. De gjør det mulig for en statistisk modell å lære og sammen med kryssvalidering, som gjør det mulig å måle den statistiske modellens ytelse. Det er et steg i riktig retning for å skape reliabilitet og validitet rundt studiet.

Stilometrisk forskning kan tidvis være påvirket av forhastelse og nødvendighet (Rudman, 1998). Disse punktene innebærer at forskning har blitt utført på datakilder som ikke var optimale eller med verktøy som ikke var tilpasset formålet. Det er etter hvert blitt utviklet et korpus (Luyckx and Daelemans, 2008b) og program (Eder et al., 2013; Juola et al., 2009) laget for stilometriske undersøkelser.

Bruk av korrumperte primære datakilder er problematisk (Rudman, 1998). Med korrumpering menes det påvirkninger datakildene kan ha vært utsatt for. Eksempelvis kan eldre tekster ha blitt påvirket av muntlige tradisjoner og dramatiske tradisjoner (Rudman, 1998). Andre farer er korrumpering via plagiering, oversettelser, imitasjon, skrivefeil, endringer i formattering og modernisering av tekst og skriving (Rudman, 1998). For eksempel kan endring av formatering utelukke bruken av enkelte applikasjonsspesifikke trekk, som avsnittslengde og oppsett.

Andre problem er når en undersøkelse feilattribuerer en forfatter, som ved CUSUM-kontroversen og svekker troverdigheten til disiplinen (Rudman, 2010). Det er viktig å poengtere at forfatteratribuering ikke gir en endelig avgjørelse om hvem om har skrevet en tekst, men sier hvem som *sannsynligvis* har skrevet en tekst. Sannsynligheten er basert på frekvenser av trekk og statistiske metoder. Det er derfor viktig å få frem at slik forskningen er idag bør stilometriske undersøkelser konkludere med omhu, særlig i rettslige og juridiske sammenhenger.

Kapittel 3

Valg av data og metode

Dette kapittelet inneholder en oversikt over metoden som er brukt til de senere undersøkelsene. Prosessen kan deles inn i tre steg: Tekstinnhenting, preprosessering av tekstene og statistisk modellering. Tekstinnhenting innebærer å lage et eget korpus. Preprosessering består av trekkutvalg og formatering av tekst. Først gis det en oversikt over potensielle tekstsamlinger som var aktuelle å bruke, deretter forklares og begrunnes valget av den endelige tekstsamlingen. Etterpå blir prosessen ved å lage et eget korpus forklart og diskutert. Etter det gis en oversikt over de andre stegene som skal utføres som en del av metoden.

3.1 Korpus

Et av de første stegene i en stilometrisk analyse er å lage eller finne en samling tekster å bruke. Hvilke tekster man velger avhenger av hva som skal undersøkes. Ved forfatterprofilering bør tekstegenskaper, som f. eks. sjanger og tema, bli kontrollert for i et korpus. Dette er for å gjøre tekstene innholdmessig homogene, for å utelukke innholdet som en påvirkende og potensielt utslagsgivende faktor i en analyse.

Stilometriske undersøkelser stiller krav til et korpus. Det er viktig at et korpus er annotert for den informasjonen man trenger. Informasjonen kan være forfatteregenskaper, sjanger-, lingvistisk eller tekstrelevant informasjon. For eksempel må man kategorisere tekster i forskjellige sjangre i forkant av en undersøkelse dersom man ønsker å undersøke forskjeller eller likheter mellom sjangere.

I denne sammenhengen har jeg valgt å undersøke om Aviskorpus ann.[1] (Hofland, 2000; Meurer, 2012) og INESS[2] (Meurer et al., 2012) er egnet til mine undersøkelser.

3.1.1 Kort om Aviskorpus ann.

Aviskorpuset ann. (AA) inneholder utvalgte materialer fra Aviskorpuset [3] som er annotert for flere variabler. Annoterte variabler i AA er blant annet lemma, ord, ordklasse, kilde, årstall, kjønn, forfatte

[1] <http://clarino.uib.no/korpuskel/page>
[2] <http://clarino.uib.no/iness/page>
[3] <http://clarino.uib.no/korpuskel/corpus-list?session-id=238532738138964>

og språk. Tekstene i AA er kategorisert etter forfattere, dato og avis. Det inneholder 35 692 210 *tokens* og 28 969 124 ord og punktum.

AA og Aviskorpus kan ikke sammenlignes direkte dersom en undersøkelse skulle brukt begge sammen, fordi de inneholder ulike tekster og tekstene ikke har individuelle identifikatorer.

AA er ikke beregnet for stilometriske undersøkelser og mangler dermed enkelte brukermuligheter. Et av disse er tekst-id eller nedlasting av tekstene. Tekst-id er viktig for å kunne kategorisere tekstene og kontrollere for tekstrelevante faktorer. En nedlastingsmulighet av tekstene hadde kunnet bøte på mangelen av tekst-id eller lagt til rette for enkelte trekk, som n-grammer.

Aviskorpus ann. ble ikke brukt i denne studien av to grunner. For det første var de mulige trekkene fra AA i stor grad leksikalske, noe jeg ikke ønsket å fokusere undersøkelsene på. Den andre grunnen var at mangelen på tekst-id og nedlastingsmulighet gjorde det vanskelig å kontrollere tekstene. Mangelen på kontroll gjorde det umulig å ekskludere uønskede tekster, lese tekstene i sin helhet og kontrollere for tema og sjanger.

AA kan brukes til stilometriske undersøkelser, med forbehold om at trekkene er begrenset, tekstene kan ikke lastes ned og forfatterne er ikke tagget presist. For eksempel kunne fremtidige studier undersøke om ordfrekvenser kan benyttes for å identifisere forskjellige norske aviser. Til mitt formål var derimot ikke AA egnet, siden målet var en studie i forfatterattribuering.

3.1.2 Kort om INESS-trebanken

Infrastructure for the Exploration of Syntax and Semantics (INESS) er en infrastruktur for forskning på trebanker, dvs. syntaktisk annoterte korpus. INESS inneholder bl.a. en relativt stor trebank for norsk med flere subkorpus. I denne studien benytter jeg subkorpusene "nob-novel_0-5", en bokmålstrebank som inneholder samlinger med noveller av forskjellige forfattere. Subkorpusene inneholder tilsammen 2 470 296 setninger og 26 903 945 ord [4].

Tekstene i INESS er parset syntaktisk med NorGram, en norsk grammatikk laget innenfor formalismen Lexical Functional Grammar (Dyvik, 2000). Korpusene er annotert for ordklasser, fraser, grammatiske kategorier og annen syntaktisk annotasjon. Gullstandarden for korpusene er annotert for hånd i INESS, resten er disambiguert stokastisk.

Målet med å bruke trebanken er å trekke ut syntaktisk informasjon fra tekstene og undersøke om denne informasjonen er nyttig i en forfatterattribueringsundersøkelse. Tidligere forskning har vist at syntaktisk informasjon kan være nyttig for å indikere forfatterskap.

Mye av denne forskningen har ikke hatt tilgang til trebanker som INESS, med grundig syntaktisk annoterte tekster. Derfor kan en undersøkelse av tekstene i INESS gi interessante og nye resultater.

For å kunne bruke subkorpusene i INESS til stilometriske undersøkelser kreves det at korpusene inneholder en viss informasjon: Forfatternavn, antall ord per tekst og at subkorpusene er annotert på samme måte. Det siste vil si at ikke alle setningene i en tekst i subkorpusene er annotert i like stor grad. Enkelte tekster har flere "uløste" setninger enn andre, det vil si setninger som ennå ikke har blitt disambiguert eller parset. Med forbehold om parsingsgrad er det mulig å lage et eget korpus.

[4] <http://clarino.uib.no/iness/page>

Tekstene i INESS er nedlastbare. Tekstene kan lastes ned og det kan hentes ut spesifikke trekk, som n-grammer. I tillegg kan trekk ekstraheres, som f. eks. frekvenser av spørresetninger, fra INESS-trebanken fra de samme tekstene. Siden de samme tekstene brukes kan resultatene også sammenlignes.

INESS ble valgt på grunn av muligheten til å kontrollere faktorer som sjanger, forfatter, størrelse, kjønn og utgivelsesår. Foruten muligheten til å kontrollere for enkelte faktorer gjør muligheten for nedlasting det mulig å sammenligne mellom enkelte trekk, n-grammer og andre INESS-spesifikke trekk.

3.1.3 Tekster til eget korpus

Ved å laste ned tekster fra forskjellige korpus i INESS laget jeg et eget korpus, "Novellekorpuset". Korpuset besto av et utvalg av 21 skjønnlitterære tekster fra 10 forskjellige forfattere. Tekstene ble tatt fra INESS-tekstsamlingene: nob_novel, nob_novel_1, nob_novel_2, nob_novel_3, nob_novel_4 og nob_novel_5 [5].

Forfatterne ble valgt ut etter hvor stor grad av tekstene som var blitt parset. Grensen ble satt til over 80 %. En høyere grense ville ikke gitt nok tekster til forsøket og en lavere grense kunne senket materialets kvalitet. Grunnen til det var et ønske om å minimere en potensiell påvirkning av frekvensene i tekster med lav parsingsgrad. For eksempel kunne det tenkes at spørresetninger ikke blir annotert i like høy grad i tekster med lav parsing. Dermed ville ikke frekvensen vært representativ for teksten, men representere mangelen av parsing. Dette kan delvis unngås ved å relativisere frekvensene med andre relative frekvenser[6].

En annen potensiell konsekvens av lav parsingsgrad er lave frekvenser. Lave frekvenser kan øke sjansen for dårlig representativitet av et trekk. For eksempel kunne en tekst med 50 % parsing ikke produsere representative frekvenser. Hvis teksten inneholdt 5 spørresetninger og 15 deklarative setninger ville det produsert en spørresetningfrekvens på 0.33 (5/15=0.33), relativt til den teksten. Dersom teksten ble videre disambiguert til 80 % kunne det påvirket annotasjonene til passivsetninger. Si at den nå inneholdt 30 spørresetninger og 350 deklarative setninger. Nå er spørresetningefrekvensen endret til 0.086 (30/350=0.086). Frekvensen er nå drastisk endret og det er rimelig å konkludere at den siste frekvensen (0.086) mer representativ for testen enn den første (0.33) når parsinggraden var på 50 %.

Tekstene ble kontrollert for om det fantes ikke-ønskelige tekster av samme forfatter i samme subkorpus. Grunnen til dette var dersom det ikke var mulig å søke etter bestemte tekster og det var nødvendig å søke etter frekvensene per forfatter. Det viste seg i ettertid at det er mulig å søke i bestemte tekster. Dermed ble ikke denne problemstillingen aktuell likevel.

Lengdene på tekstene varierer mellom 4508 ord og 100 563 ord. Alle forfatterne har to tekster i korpuset, unntatt en forfatter som har tre. Den forfatteren fikk tre tekster siden alle var parset på over 80 % og lå i samme korpus. En liste over tekstene kan sees i tabell 3.1 på s. 26.

[5] <http://clarino.uib.no/iness/page>

[6] Relativisering av frekvenser blir diskutert videre i kapittel 5

	Forfatter	Tittel	Utgitt	Original	Sjanger
ar1	Ragde, Anne B.	En kald dag i helvete	1999	Norsk	Roman
ar2	Ragde, Anne B.	Ansiktet som solen	1996	Norsk	Kortprosa
mj1	Johnsgaard, Magnar	Vanskapningen	1991	Norsk	Roman
mj2	Johnsgaard, Magnar	Veivokteren	1992	Norsk	Kortprosa
ph1	Hagan, Patricia	Vill og vakker	1991	Engelsk	Roman
ph2	Hagan, Patricia	Kjærlighetens triumf	1991	Engelsk	Roman
ul1	Lindell, Unni	Slangebæreren	1996	Norsk	Roman
ul2	Lindell, Unni	En grusom kvinnes bekjennelser	1993	Norsk	Kortprosa
ao1	Oterholm, Anne	Avbrutt selskap	1996	Norsk	Roman
ao2	Oterholm, Anne	Avslutningen	1999	Norsk	Roman
ok1	Klippenvåg, Odd	Body and Soul	1998	Norsk	Kortprosa
ok2	Klippenvåg, Odd	Bruckner, en lengsel	1997	Norsk	Roman
ok3	Klippenvåg, Odd	Et virkelig liv	1991	Norsk	Roman
el1	Lindvåg, Ellen Iris	Ingen kan nå med	1998	Norsk	Roman
el2	Lindvåg, Ellen Iris	Sett: ?	1997	Norsk	Roman
re1	Enger, Rolf	Solformørkelse	1994	Norsk	Roman
re2	Enger, Rolf	Hvis noen skulle være så vemmelige	1997	Norsk	Kortprosa
dh1	Hamilton, Dan	Kameleonkvinnen	1994	Engelsk	Roman
dh2	Hamilton, Dan	Tiggerkongen	1993	Engelsk	Roman
ek1	Kiøsterud, Erland	Ved fjellets fot	1991	Norsk	Kortprosa
ek2	Kiøsterud, Erland	Den norske sangeren	1999	Norsk	Kortprosa

Tabell 3.1: Oversikt over tekstene i "Novellekorpuset"

Tekstene er prosa bestående av noveller og romaner. 17 hadde norsk som originalspråk og 4 hadde engelsk som originalspråk. Tekstene ble utgitt i tidsperioden 1991-1999.

Alle tekstene hadde en lesbarhetsindeksering som var "lettlest" i følge lesbahetsindeksen, Liks[7]. Liks måler lesbarheten til en tekst med en formel, utviklet av Björnsson (1968). Formelen ble utviklet og testet på svenske tekster, men kan også brukes på norske tekster[8]. Lesbarheten til tekster grupperes i en skala fra 1 til 60 [9]. 1 er det mest lettskrevne. Liks-formelen (Björnsson, 1968, s. 66 og s. 214):

$$\frac{O}{S} + \frac{Lo \times 100}{O} = LIX^{10}$$

O = Totalt antall ord i tekst

S = Totalt antall setninger

[7]Kalkulator: <http://www.lix.se> drevet G. Seimyr ved <http://www.semios.se>

[8]<www.sprakradet.no/Klarsprak/sprakhjelp/Skriverad/Om-liksberegning/> og <http://http://www.skriftlig.no/liks-kalkulator/>.

[9]Liks kan gå høyere enn 60, men den høyeste vanskelighetsgraden begynner på 60 og det er derfor brukt som stoppunkt her.

[10]LIX er den svenske betegnelsen på Liks.

	Setninger	Ord	Liks	GSL	TTR	OVIX	OVR
ar1	5560	53511	23	8.40	11.68	61.14	80.39
ar2	3692	38446	24	9.33	15.80	66.28	82.62
mj1	5451	44370	19	7.29	11.25	58.38	79.73
mj2	4208	34254	22	8.18	13.02	59.58	80.65
ph1	7494	80692	26	11.16	9.90	61.59	79.69
ph2	6592	73836	27	11.70	9.92	60.69	79.54
ul1	11285	100563	24	7.73	9.23	61.90	79.43
ul2	5501	46779	23	7.96	12.79	62.35	81.02
ao1	4130	27082	19	6.56	8.97	49.14	76.64
ao2	6256	38477	18	6.35	8.58	51.62	77.01
ok1	3742	33984	24	9.32	14.88	63.41	81.92
ok2	2729	25337	26	9.45	17.93	66.02	83.22
ok3	7287	79952	25	9.44	11.68	65.38	81.06
el1	7371	50006	19	6.13	11.98	61.36	80.55
el2	9573	68070	21	8.06	11.89	65.00	81.08
re1	4741	43687	24	9.87	14.56	65.64	82.14
re2	3903	39901	25	11.00	16.44	68.45	83.10
dh1	3261	32088	26	10.46	14.23	61.34	81.38
dh2	3390	33377	25	10.74	13.58	60.46	81.01
ek1	611	4508	27	8.42	26.05	58.38	84.25
ek2	1724	17821	30	11.20	21.04	67.43	84.22

Tabell 3.2: Egenskaper av innhold i korpuset

Lo = Antall lange ord (lenger enn 6 bokstaver)

Tekstene i korpuset hadde verdier mellom 18 og 30 på Liks-skalaen. Under 30 er definert som "lettsleste" til "svært lettleste" bøker og 30-40 på skalaen er definert som "lettleste" til "middelmådige" bøker (Björnsson, 1968, s. 89).

Oppsettet på de nedlastede tekstene tilsvarer ikke originaltekstene. I de nedlastede filene er hver setning plassert på en egen linje. Dette utelukker enkelte applikasjonsspesifikke trekk, f. eks avsnittslengde.

3.1.4 Korpusets egenskaper

En svakhet ved korpuset er at det ble bygget på antagelsen om at det ikke var mulig å søke etter bestemte tekster i et korpus i INESS, men at man måtte søke ut fra forfattere. Det førte til at enkelte forfattere ble utelukket fordi de inneholdt ønskede og uønskede tekster i subkorpusene. De uønskede tekstene med lav parsingsgrad ble antatt å ikke kunne ekskluderes. Dokumentasjonen til trebanken manglet informasjon om søke ut ifra tekst[11]. Dette førte til at tekster som kunne bli inkludert ikke ble det og en forfatter fikk 3 tekster med i korpuset i motsetning til 2 tekster. En annen konsekvens

[11] <http://clarino.uib.no/iness/page?page-id=iness-documentation>

var at tekstene ikke kunne bli kontrollert for tekstlengde i denne delen av prosessen, noe som førte til at enkelte forfattere har mye større tekstmengde enn andre.

Tekstlengde kan kontrolleres for i de nedlastede filene. Tekstlengde kan kontrolleres for i preprosesseringsdelen, ved å kutte tekstene i mindre deler eller ved å bruke relative frekvenser. Et eksempel på relativ frekvens er frekvensen av ordet "humle" i en tekst. Hvis antall forekomst av "humle" blir delt på antall ord i teksten ordene kommer fra er dette en relativ frekvens.

Den ene forfatteren har 3 tekster i stedet for 2, som de andre forfatterne har. Dette kan være problematisk, men tekstene har ikke mer enn 140 000 ord til sammen[12]. At Klippenvåg har tre forskjellige tekster *kan* gjøre at det er større varians i forsøket på å fange opp hans skrivestil eller omvendt. Hvis et forsøk gir frekvensene 1, 3, 3 per tekst er det noe annet enn hvis frekvensene er 3, 3, 6. I det første tilfellet ville en modell hatt lettere for å se likhetene mellom forfatterens tekst dersom det siste tallet er med. I det andre tilfellet er det siste tallet med på å gjøre frekvensene til forfatteren mer ulik.

En annen svakhet er at enkelte av tekstene ikke er kontrollert for tema eller sjanger. Alle tekstene er prosatekster, både romaner og novellesamlinger. Foruten varierer det hvilke typer romaner og noveller som er med. For eksempel er "Slangebæreren" en kriminalroman og "Kjærlighetens triumf" en kjærlighetsroman. Det er heller ikke kontrollert for at forfatterne skal ha en variasjon innen tekstutvalget. Det betyr at forfattere som Hagan har to tekster som begge er kjærlighetsromaner. Dette kan være uheldig og det kan føre til type-1-feil, altså falske positive funn.

4 av tekstene er oversatt fra engelsk, noe som ble oppdaget i ettertid. Det kan vise seg å være interessant å undersøke om dette kan detekteres i forsøkene. Det kan være problematisk dersom tekstene har hatt forskjellige oversettere kan dette ha påvirket oversettelsene av tekstene ved ord- og setningsvalg. En annen faktor kan være at originalspråket til teksten påvirker oversettelsen på en målbar måte. På en annen side kan dette også sees på som en egenskap ved forfatterens skrivestil, men dette er noe det kan tas forbehold om.

En av hovedstyrkene til korpuset er at alle forfatterne har lange tekster. Jo større tekstsamling en forfatter har, jo lettere er det å gjenkjenne skrivestilen. Det er fordi lange tekster gir økt mulighet for å måle de ønskede frekvensene. Det betyr at i et tilfelle hvor man hadde 100 ord lange e-poster fra en forfatter og 100 ord lange dikt fra en annen forfatter hadde disse forfatterne trolig blitt kategorisert riktig, men potensielt på grunn av de særegne formene til tekstene. Dersom en hadde at forfattere med hver sin e-post og dikt på 100 ord er det mindre sjanse for at en statistisk modell hadde kunnet kategorisert de riktig. Biber (1990, 1993b) forslår minst 1000 ord lange tekster for oppnå representativitet. Den korteste teksten i dette korpuset er på 4500 ord.

Det finnes ingen konsensus over hvor mange forfattere et forfatterattribueringsforsøk bør inneholde, men jeg mener 10 forfattere er tilstrekkelig til undersøkelsene mine. Færre forfattere og tekster kunne gjort resultatene vanskeligere å generalisere og tolke.

Videre har hver av forfatterne mer en en tekst i korpuset, for å minimere sjansen for at utslaget ikke er *tekstavhengig*, men *forfatteravhengig*. Forfatterne er også kontrollert for kjønn med 5 kvinnelige og 5 mannlige forfattere hver og tidsperiode (1990-tallet). Dette er en fordel siden det

[12]I motsetning til Hagan som har en ordmengde på over 154 000 til sammen

minimerer språklig variasjon som blir påvirket av tid. Dette kan være setningsbruk, ordbruk og bokmålsformer.

3.1.5 Korpusets egenskaper i sammenheng med trekk fra søk i INESS

Et forbehold som må tas ved bruken av tekstene, er parsingsgraden av tekstene. Siden parsingsgraden er 80 % og over er det mulig at tekstene ikke har tilsvarende parsingsgrad og dermed tilsvarende og sammenlignbar frekvens. For eksempel kan det tenkes at ved 80 % diambigueringsgrad er spørresetninger annotert i 60 % av tilfellene, men med 95 % parsingsgrad er spørresetninger annotert i 90 % av tilfellene.

En måte å unngå denne problemstillingen på er å søke på syntaktiske kontruksjoner som er lett disambiguerbare eller høyfrekvente i INESS.

Et annet punkt ved bruken INESS er dersom setningene i en tekst er disambiguert for hånd. INESS har forskjellige lingvister som disambiguerer setninger. Hvis noen av lingvistene tolker og disambiguerer setninger ulikt, f. eks. flertydige setninger, kan dette gi utslag i frekvensene man får fram i INESS.

Det er et forhold mellom lesbarhetsindeksen og graden av parsing i INESS som man må være bevisst på. Alle tekstene jeg har lagt i korpuset ligger mellom 18-30 på lesbarhetsindeksen og har en høy parsingsgrad i INESS. Dette forholdet kan ha sammenheng med at lettleste tekster enklere blir disambiguert i INESS.

Bruken av Liks for å fastslå hvor lesbar en tekst er i seg selv problematisk. Liks tar for eksempel ikke hensyn til sjeldne ord eller krevende syntaks. Dette betyr at undersøkelser med Liks må tas med forbehold om at det er en overflatisk og begrenset måte å kvantifisere lesbarheten av en tekst.

3.2 Preprosessering av tekster

Preprosessering av tekster er neste steg etter valget eller dannelsen av et korpus. I denne delen blir tekstene omformatert for å kunne sendes gjennom en statistisk modell eller ekstrahere trekk.

En preprosessering krever at man først vet hvilke lingvistiske trekk man ønsker å undersøke og hvilke programmer og algoritmer som skal brukes. Dette er begrenset av hvordan korpuset man bruker er annotert siden man ikke kan undersøke noe man ikke har eller kan annotere dette korpuset for. Når man vet hvilke trekk man ønsker å undersøke, kan man deretter formatere tekstene i korpuset slik at de kan sendes gjennom den modellen og det programmet man har valgt. Eksempelvis er enkelte læringsalgoritmer bedre enn andre til å håndtere ikke-numeriske verdier og andre håndterer numeriske verdier bedre.

WEKA krever spesifikke filformat, formatering av tekstene og at filene har en *header*. *Stylo* krever ikke formatering av tekstene, men krever en kategorisering som legges i tittelen på tekstens filnavn.

3.3 Statistiske og datamaskinelle modeller

Modellene ble valgt etter hvilke som egnet seg til de spesifikke datasettene og hvilke som var å finne i programmene. For eksempel ble primært ikke-overvåkede algortimer valgt i *Stylo* fordi programmet primært er rettet mot denne typen statistikk.

For ord- og tegnbaserte trekk ble *Stylo*-programmet brukt. Av statistiske metoder ble, *multidimensional scaling*, *cluster analysis* og *consensus tree* brukt. Sammen med en av metodene ble distansen satt til *classic delta*.

For de syntaktiske trekkene ble WEKA brukt, med overvåkede læringsmetoder. *k-Nearest Neighbour*, *SVM*, *Naive Bayes* og *decision trees* ble brukt i forsøk utført med trekkene. De ble valgt fordi de egnet seg til numeriske fremstillinger av trekkene, både kontinuerlige og diskrete fremstillinger.

For de syntaktiske trekkene kunne andre programmer bli brukt, f. eks. TiMBL [13]. TiMBL ble valgt bort fordi programmet hadde færre valgmuligheter av algoritmer og håndterte ikke numeriske verdier automatisk.

3.4 Sammenhenger mellom korpus-tekstene

En videre undersøkelse ble gjort av korpuset. Dette ble gjort for å undersøke om tekstene hadde underliggende likheter, som kunne påvirke resultatene av forsøkene senere.

3.4.1 Undersøkelse av sammenheng mellom lesbarhetsindeks og parsingsgrad

For å avdekke om det var en sammenheng mellom lesbarhesindeksen og parsingsgraden av tekstene fra INESS, tok jeg utgangspunkt i korpuset "nob_novel", siden det inneholdt tekster jeg hadde brukt. Jeg kunne brukt et eller flere av de 5 andre korpusene som jeg tidligere hadde benyttet, men jeg mente "nob_novel" var tilstrekkelig til denne undersøkelsen. Korpuset "nob_novel" inneholdt 45 tekster med varierende parsingsgrad.

En tabell ble generert med oversikt over tekstene i subkorpuset og graden av parsing. Etterpå ble annenhver tekst lastet ned, slik at jeg fikk 23 tekster å sammenligne med. Tekstene fikk kalkulert lesbahetsindeksen sin, som ble lagt inn i den nedlastede tabellen.

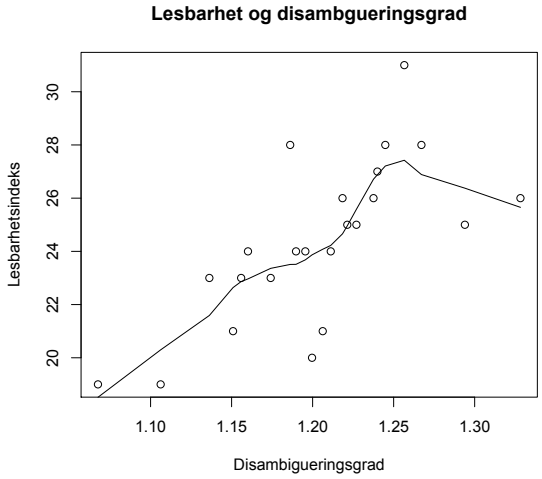
Etterpå ble tabellen omgjort til en .csv-fil og lagt inn i statistikkprogrammet R. I R ble antall disambiguerte setninger i en tekst delt på antall setninger som ikke var disambiguert, enten delvis eller fullstendig. Etter å fått bekreftet at variablene hadde et lineært forhold og bestemt graden av outliers ved å plassere variablene i en graf og utførte en Shapiro-test på hver variabel for å sjekke normalitet, utførte jeg en Pearsons-korrelasjonstest. Korrelasjonen var på 0.70, som kan anses som en moderat positiv korrelasjon.

Dette betyr at korpuset jeg har satt sammen har blitt influert av lesbarhetsgraden av tekstene. Tekstene i korpuset mitt er mer lettleste enn de ellers ville vært dersom jeg ikke hadde graden av

[13] <http://ilk.uvt.nl/timbl/>

parsing som et kriterium for valget av tekster fra INESS.

Figur 3.1: Sammenheng mellom lesbarhet og parsing



3.4.2 Undersøkelse av andre sammenhenger mellom tekstene

Tekstene i korpuset med variablene, jfr. tabellene 3.1 på s. 26 og tabell 3.2 på s. 27 ble deretter testet i R med Pearsons korrelasjonstest. Først ble .csv-filen med tekstene lest inn i R, deretter ble variablene brukt i korrelasjonstesten.

Eksempel:

```
cor.test{dat\$OVIX,dat\$OVR)
```

Forkortelsene på noen av begrepene:

- *Lesbarhetsindeks* = Liks
- *Gjennomsnittlig setningslengde* = GSL
- *Type/token ratio* = TTR
- *Ordvariasjonsindeksen* = OVIX
- *Ordvariasjonsratio* = OVR

OVIX er regnet ut med formelen og tar bedre hensyn til tekstlengde enn TTR[14]:

$$\frac{log(tokens)}{log(2 - (log(types)/log(tokens)))}$$

[14] <http://www.lix.se/index.php>

OVR regnes ut med formelen og er bedre enn TTR med hensyn til tekstlengde[15]:

$$\frac{log(types)}{log(tokens)}$$

Tabell 3.3: Sammenhenger innad tekstene

Sammenheng	Korrelasjon
OVR/TTR	0.88
OVIX/OVR	0.78
Ord/Setninger	0.92
GSL/Liks	0.85
GSL/OVIX	0.59
Ord/Liks	-0.08

Det er ikke overraskende at OVR og TTR korrelerer, slik man ser i tabell 3.4.2. Begge er basert på beregninger på variasjonen av ord i forhold til størrelsen på teksten. At OVIX OG OVR også korrelerer er av samme grunn naturlig, da begge er basert på ordvariasjon.

Antall ord og antall setninger korrelerer sterkt. Trolig fordi antall ord i en tekst er sterkt avhengig av antall setninger og omvendt.

GSL og Liks korrelerer også svært sterkt. Trolig fordi deler av Liks-formelen er basert på GSL.

GSL og OVIX korrelerer moderat. GSL og OVIX er ikke basert på hverandre, men begge er basert på ord. GSL er basert på antall ord per setning og OVIX uttrykker en form for ordvariasjon.

Antall ord og Liks korrelerer ikke. Trolig fordi størrelsen på en tekst ikke nødvendigvis sier noe som hvor lett eller vanskelig den er å lese. Tekstutvalget her er antageligvis ikke en god indikasjon på korrelasjon mellom ord og Liks generelt.

3.5 Oversikt over stilometriske forsøk

I de neste kapitlene vil jeg beskrive undersøkelsene og resultatene i mer detalj. Tabell 3.4 er ment å gi en oversikt over undersøkelsene er gjort og er beskrevet i de neste kapitlene. Undersøkelsene som har brukt *Stylo* finns i kapittel 4 og undersøkelsene som har brukt WEKA finns i kapittel 5.

Kort kan det sies at undersøkelsene utført ved hjelp av *Stylo* har tatt for seg de selvlagde korpusene og brukt ikke-overvåkede og overvåkede læringsalgoritmer. Fremstillingen av resultatene er hovedsakelig grafisk. Undersøkelsene utført i WEKA har brukt frekvenser hentet fra INESS. Frekvensene kommer fra forskjellige morfo-syntaktiske og syntaktiske trekk i de samme tekstene som ligger i det selvlagde korpuset.

[15] <http://www.lix.se/index.php>

Tabell 3.4: Oversikt over forsøk

Algoritme	Program	Korpus	Datatype	Fremstilling
Consensus Tree	Stylo	Novellekorpus	Ikke-numerisk	Grafisk
Cluster Analysis	Stylo	Novellekorpus	Ikke-numerisk	Grafisk
Multidimensional Scaling	Stylo	Novellekorpus	Ikke-numerisk	Grafisk
Cluster Analysis	Stylo	Kjønnskorpus	Ikke-numerisk	Grafisk
Consensus Tree	Stylo	Likelangtkorpus	Ikke-numerisk	Grafisk
Cluster Analysis	Stylo	Likelangtkorpus	Ikke-numerisk	Grafisk
k-NN	WEKA	INESS	Diskret	Tabell
Naive Bayes	WEKA	INESS	Diskret	Tabell
SVM	WEKA	INESS	Diskret/Kontinuerlig	Tabell
Decision tree[16]	WEKA	INESS	Diskret/Kontinuerlig	Tabell

Kapittel 4

Stylo - trekk og forsøk

4.1 Om Stylo og valgmuligheter i programmet

Stylo (Versjon 0.5.8-2) er en pakke til R utviklet for stilometriske undersøkelser (Eder et al., 2013). Den inneholder en rekke verktøy for å gjøre grunnleggende stilometriske forsøk, både ikke-overvåket og med overvåket læring. For ikke-overvåket læring brukes *stylo()* -funksjonen. *Stylo()* kan anvende *Principal Components Analysis*, *Cluster Analysis*, *Multidimensional Scaling*, and *Bootstrap Consensus Trees* (Eder et al., 2013). For å bruke overvåket læring kan *classify()*-funksjonen brukes. *classify()* inneholder mulighet for SVM, *k-Nearest Neighbors*, *Naive Bayes*, *Nearest Shrunken Centroids* (NSC) (Jockers et al., 2008) og Delta (Burrows, 2002).

Statistiske distanser kan velges i pakken. Det er mulig å velge mellom *Classic Delta*, *Argamon's Delta*, *Eder's Delta*, *Eder's Simple*, *Manhatten*, *Canberra* og evklidisk distanse. I tillegg finnes det flere andre skript i pakken. Ett av dem *Rolling Delta* som analyserer tekster skrevet av flere forfatter og forsøker å skille de ulike delene skrevet av hver forfatter fra hverandre.

Stylo har innstillinger for noen språk, blant annet tysk og engelsk. For eksempel, for å kunne fjerne pronomenene i teksten fra å bli analysert. Iflølge en studie som fjernet pronomen økte nøyaktigheten av Delta-testing av tekster (Hoover, 2004b). Norsk er ikke et inkludert språk i *Stylo*. Det gjør at pronomene i en tekst ikke kan fjernes automatisk, som er vist å forbedre resultatene i forfatteratribueringsundersøkelser i enkelte språk (Eder et al., 2014).

Stylo henter filene i mapper og analyserer alle samtidig i filformatene, .txt, .xml og .html.

Hovedtrekkene i pakken er n-gram, enten bokstavgram eller ordgram. En valgmulighet for å manipulere trekkene er bruken av *Most Frequent Words*[1] (MFW). Med MFW menes det at en liste laget med de mest frekvente ordene per tekst. Frekvensene blir deretter normalisert for å lage en endelig ordliste for den statistiske analysen. Dersom *consensus tree* blir valgt viser resultatet et kompromiss mellom de underliggende *cluster* analysene. Man kan stille inn listen på minimum, maksimum og økning. Økninginnstillingen blir brukt dersom man gjør en *consensus tree* analyse. *Culling* i *Stylo* referer til en manipulering av ordlisten (Hoover, 2004b,a). *Culling*-innstillingene spesifiserer i hvilken grad ord eller n-grammer som ikke forekommer i alle tekstene som undersøkes

[1] Eller i n-gram med bokstaver blir dette *Most Frequent Characters* (MFC), men for enkelthetens skyld kommer jeg kun til å bruke forkortelsen MFW, men i enkelte R-genererte tabeller vil forkortelsen MFC forekomme i diagrammene.

skal fjernes (Eder et al., 2013). En verdi på 20 vil si at ordene forekommer i minst 20 % av tekstene vil bli tatt med i analysen. I likhet med MFW finnes det innstillinger for minimum, maksimum og økning.

Stylo kan også ta høyde for lister av eksisterende frekvenser, ordlister og skille mellom stor og liten bokstav. Det er også mulig å dele teksten opp i mindre deler eller ta tilfeldige deler av dem for å analysere dem. Dette kalles *sampling*. En annen innstilling er *list cutoff* som kutter frekvenslisten på et gitt punkt, for eksempel etter 1000 ord. Det er for å gjøre enkelte krevende forsøk mindre datamaskinelt krevende.

Resultatene kan skrives ut i forskjellige typer grafer og formater, som trediagrammer eller prikkdiagrammer.

Dersom ingenting er spesifisert i de kommende forsøkene er innstillingene: Språk = *other*, *encoding* = utf-8, ingen *sampling*, *list cutoff* = 5000, ingen skille mellom stor/liten bokstav og ingen bruk av eksisterende frekvens- eller andre lister. MFW- og *culling*-instillingene blir spesifisert som for eksempel MFW = 0/10/5, som her betyr: 0 = minimum, 10 = maksimum og 5 = økning.

Distansen jeg har valgt er *Classic Delta*, som appliseres til normaliserte ordfrekvenser (Eder et al., 2013). *Classic Delta* er ikke sensitiv for antall tekster i korpuset, men distansen er sensitiv for overvekt av en klasse. Eksempelvis om korpuset inneholder 10 tekster av en forfatter og 1 tekst av en annen forfatter. Distansen er også egnet til engelsk, og det antas her at likhetene i morfologi sammenfaller tilstrekkelig til at den egner seg til norske tekster. *Classic Delta* kan kontrasteres til *Eder's Delta*, som er laget til språk med høy grad av morfologisk bøyning.

4.2 Preprosessering av tekstene til Stylo

Tekstene ble lastet ned fra INESS-trebanken og lagt i en egen mappe. Tekstene ble lagret som .txt-filer og fikk navn først etter klasse, deretter understrek og til sist navnet på prosateksten.

Eksempel:

```
``Johnsgaard_Veivokteren.txt"
```

Korpuset fikk navnet "Novellekorpuset", for å skille det fra andre korpus i preprosesseringsfasen. Foruten å bestemme klasser, mappe og lengde på tekstene krevdes det ikke mer preprosessering.

Tabell 4.1: Oversikt over korpus til Stylo

Korpusnavn	Størrelse	Kategorier
Novellekorpus	Hele tekster	Forfatter
Likelangtkorpus	4500 ord	Forfatter
Kjønnskorpus	4500 ord	Kjønn

4.2.1 Andre korpus - "Likelangtkorpuset" og "Kjønnskorpuset"

Fordi tekstlengden varierte innad i korpuset laget jeg et subkorpus med de samme tekstene. I dette korpuset ble tekstene kuttet ned til 4500 ord hver. 4500 ord ble valgt fordi det var lengden på den korteste teksten.

Korpuset ble laget ved hjelp av skriptet wordsplitter.sh, se vedlegg B s. 89. Skriptet legger hvert enkelt ord på en ny linje og kutter teksten på linje nr. 4500. Til slutt legges ordene tilbake i setninger. Den eneste endringen fra tidligere er at teksten ikke begynner på ny linje ved ny setning.

I og med at dette var en måte INESS valgte å fremstille tekstene på og ikke forfatterne selv og denne typen applikasjonsspesifikke trekk uansett ikke skulle brukes, tror jeg ikke oppsettet vil påvirke resultatene. Korpuset ble kalt "Likelangtkorpuset" for å skille det fra de andre korpusene i teksten.

Ut i fra "Likelangtkorpust" ble et nytt korpus laget, "Kjønnskorpuset", hvor tekstene ble sortert etter forfatterens kjønn. Det ble laget ved å legge en kopi av tekstene i en ny mappe og endre navne til å legge den nye klassifiseringen først i navnet,K for kvinner og M for menn. For eksempel ble teksten "Slangebæreren" av Unni Lindell gitt navnet,

```
K_Lindell_Slangebæreren.txt
```

i stedet for benevnelsene brukt i de andre korpusene:

```
``Lindell_Slangebæreren.txt"
```

I de andre korpusene er forfatternavnet klassifiseringen.

Tekstene fra alle korpusene ble lastet ned og lagt i egne mapper for hvert korpus. Tekstene ble lagret etter forfatternavn og navn på tekst, som beskrevet for "Kjønnskorpuset" ovenfor. *Stylo* krever at klassifiseringsnavnet til en tekst blir lagret i filnavnet.

4.2.2 Preprosessering i R

```
> library(stylo)
> getwd()
> setwd("/Users/victoriatroland/Desktop/INESS korpus/Stylo-forsøk/
Kjønnskorpus")
> stylo(corpus.dir="Folder name", gui=TRUE)
```

Forsøkene i R starter med de ovenstående linjene[2]. *gui=TRUE* gir en grafisk fremstilling av *Stylo*. Resten av forsøket blir gjort i det grafiske vinduet. Valgene blir nærmere presisert i forsøkene.

4.3 Forsøk med "Likelangtkorpuset"

"Likelangtkorpuset" består av 21 tekster på hver 4500 ord av 10 forskjellige forfattere.

[2] "Kjønnskorpus" ble erstattet med "Novellekorpus" eller "Likelangtkorpus" etter behov.

Tabell 4.2: Oversikt over forsøk i Likelangtkorpus

Forsøk nr.	Algoritme	Distanse	n-gram	MFW	Culling
1.1	Consensus Tree	Classic Delta	3-bokstavgram	0/1000/100	0/0/0
1.2	Consensus Tree	Classic Delta	2-ordgram	0/1000/100	0/0/0
2.1	Consensus Tree	Classic Delta	2-ordgram	0/0/0	0/100/20
2.2	Cluster Analysis	Classic Delta	2-ordgram	0/0/0	0/0/0
2.3	Cluster Analysis	Classic Delta	2-ordgram	0/1000/100	0/100/20
2.4	Cluster Analysis	Classic Delta	2-ordgram	0/0/0	0/100/20
2.5	Cluster Analysis	Classic Delta	2-ordgram	0/1000/100	0/0/0

4.3.1 "Bootstrap Consensus Tree"-forsøk

Det første forsøket med dette korpuset var et forsøk med økning av MFW og med bruk av *consensus trees* med *bootstrapping*. Motivet er å undersøke hvordan en endring i MFW påvirker hvor godt modellen forutsier forfatterskap.

Forsøk 1.1 parametrer: n-gram=3 bokstaver, MFW=0/1000/100, culling=0/0/0, consensus tree=0.5 og classic delta. Se C.1 i C på s. 91.

Forsøk 1.2 parameter: n-gram=2 ord, MFW=0/1000/100, culling=0/0/0, consensus tree=0.5 og classic delta. Se C.2 i C på s. 92.

Dendrogrammene fra forsøket viser en lovende klassifisering med begge trekkene som ble valgt i forsøket. Jevnt over samsvarte begge trekke og resultatene med hverandre. Resultatene blir beskrevet nærmere i 4.3.3 på s. 39.

4.3.2 Parametrene MFW og *culling* gjensidige påvirkning

I dette forsøket ble det gjort forsøk med og uten MFW og *culling*. Motivet var å undersøke hvordan parametrene påvriket resultatene og hverandre.

Forsøk 2.1 parametrer: n-gram=2 ord, MFW=0/0/0, culling=0/100/20, consensus tree=0.5 og classic delta. Se C.3 i C på s. 93.

Forsøk 2.2 parametrer: n-gram=2 ord, MFW=0/0/0, culling=0/0/0, cluster analysis og classic delta. Se C.4 i C på s. 94.

Forsøk 2.3 parametrer: n-gram=2 ord, MFW=0/1000/100, culling=0/100/20, cluser analysis og classic delta. Se C.5 i C på s. 95.

Forsøk 2.4 parametrer: n-gram=2 ord, MFW=0/0/0, culling=0/100/20, cluster analysis og classic delta. Se C.6 i C på s. 96.

Forsøk 2.5 parametrer: n-gram=2 ord, MFW=0/1000/100, culling=0/0/0, cluster analysis og classic delta. Se C.7 i C på s. 97.

4.3.3 Funn med "Likelangtkorpuset"

Generelt var det store likheter på tvers av typene n-gram: Bokstavgrammer og ordgrammer. De dårligste resultatene forekom når *culling* ble brukt uten MFW og uten hverken MFW eller *culling*.

De mest presise forfatterkategoriseringene forekom når MFW ble brukt, med eller uten *culling*. Dette indikerer at manipulering av ordlisten kan være svært effektivt med n-gram.

Bootstrap Consensus Tree-forsøk

I 4.3.1 ble det utført et forsøk med økning av MFW og *consensus tree*. Resultatene illustreres av figurene C.1 og C.2 i C på s. 91. Grafene til forsøket med *bootstap consensus*-metoden. I den første grafen ble bokstav-gram brukt med MFW=0-1000.

Grafene viser at bruk av parametrene *consensus tree*, MFW og bokstav-gram kan skille de 10 ulike forfatterne fra hverandre. Den største feilklassifiseringen var Klippenvåg, som ble klassifisert for seg selv. Enger sine tekster er plassert nært hverandre, men ikke på samme gren. Hagan og Hamilton sine tekster er klassifiserte på samme hovedgren. Tekstene er også de eneste som er oversatt fra engelsk og det kan tenkes at det er en sammenheng her.

I en lignende undersøkelse med ord-gram (n=2) (se s. 92 i C), ble ikke tekstene til Enger klassifisert sammen. I denne undersøkelsen ble Johnsgaard sine tekster feilklassifisert og plassert ved siden av hverandre.

Culling og MFW-parametrene i forsøk

En forfatter som skiller seg ut er Oterholm. Oterholm blir i alle forsøkene kategorisert riktig, se C.2 på s. 93. En annen forfatter er Lindell som også i stor grad kategorisert riktig, men ofte nærmere andre forfattere. Alle forsøkene er gjort med ordgram (n=2) og parametrene varierte med eller uten MFW og *culling*. Dette sier noe om hvor distinkt frekvensene til både Oterholm og Lindell er i forhold til de andre forfatterne.

I forsøkene med *culling* mot ikke-*culling*, var MFW satt til 0. Se figur C.3 (*consensus tree*) på s. 93 og C.6 (*cluster analysis*) på s. 96 for *culling* og figur C.4 på s. 94 (*cluster analysis*) for ikke *culling*.

De mest påfallende med forsøkene, bortsett fra Oterholm og Lindell konsekvent ble riktig gruppert, var jevnt over stor grad av feilgruppering av de andre forfatterne, når MFW = 0. Hverken med eller uten *culling*-parameteret grupperes forfatterne påfallende bedre. Når et forsøk har MFW=0-1000 blir forfatterne gruppert riktig i større grad, men dette gjelder både med og uten *culling*. Ut fra disse forsøkene var ikke *culling* et nyttig parameter for å gruppere forfatterskap.

MFW-forsøkene hvor det ble utført forsøk med og uten MFW inneholdt større variasjon innad grupperingsgraden. Forsøkene med MFW=1000 (se figurene C.5 (med *culling*) og C.7 (uten *culling*) på s. 95 og 97) er svært like. I begge er kun Enger feilgruppert, og en av tekstene ("Hvis noen kunne være så vemmelige") er plassert i nærheten av Lindvåg og den andre teksten ("Solformørkelse") gruppert i nærheten av Lindell. Sammenlignet med forsøket uten MFW (uten *culling*) (se figur C.4 på s. 94) er forsøket med MFW i mye større grad riktig gruppert. Med MFW=0 er kun

Oterholm og Lindell riktig gruppert, i motsetning til MFW=1000 hvor kun Enger er feilgruppert. MFW-parameteret kan tydeligvis ha en stor innvirkning på grupperingen av forfatterskap, i alle fall når det er satt til 0/1000/100 (med 100 ord økning for hver gjennomgang av forsøket).

4.4 Forsøk og funn med ”Novellekorpus”

“Novellekorpuset” består av 21 tekster av 10 forfattere med original tekstlengde. Formålet med å bruke dette korpuset i tillegg til “Likelangtkorpuset” er å sammenligne resultatene i tilsvarende statistiske undersøkelser. I tillegg kan “Novellekorpuset” være interessant for å utforske *sampling*-innstillingene i *Stylo*. *Sampling* er valget for å kutte tekstene i like store deler eller ta et tilfeldig utdrag av tekstene basert på en forutbestemt lengde.

Sampling i *Stylo* kan være en interessant måte for å se om alle delene av teksten til en forfatter ligger nært hverandre. Ved å dele teksten opp med *sampling* kan vi undersøke dette. Ved å bruke *random sampling* er det mulig å undersøke om tilfeldige utdrag av teksten også er en måte som kan attribuere en forfatter.

Parameter i forsøk 4.1, 4.2 og 4.3: *char*= 3, *list cutoff* = 3000, MFW=1000/1000/0, MDS, *classic delta*, *sampling size* = 4000.

Variasjonene er: 4.1. ingen *sampling*, 4.2. *normal sampling*og 4.3. *random sampling*.

Tabell 4.3: Oversikt over forsøk i Novellekorpus

Forsøk nr.	Algoritme	Distanse	n-gram	MFW	Culling	Sampling
3.1	Consensus Tree	Classic Delta	2-ordgram	0/1000/100	0/100/20	Nei
3.2	Consensus Tree	Classic Delta	2-ordgram	0/0/0	0/100/20	Nei
4.1	MDS	Classic Delta	3-bokstavgram	0/1000/100	0/0/0	Normal
4.2	MDS	Classic Delta	3-bokstavgram	0/1000/100	0/0/0	Nei
4.3	MDS	Classic Delta	3-bokstavgram	0/1000/100	0/0/0	Random

4.4.1 Forsøk mellom “Novellekorpus” og “Likelangtkorpus”

Ved å utføre en sammenligning av to forsøk med like parametre, med to forskjellige tekstlengder i korpus var det mulig å undersøke om forskjeller med tekstlengde var målbart. Fordelen med å bruke *bootstrap consensus tree”* var gjennomgangene med forskjellige parametre. Gjennomgangene blir til sist sammenlignet og plassert i et *consensus tree*. Sammenligningen kan gi en indikasjon om innvirkningen på tekstlengde i korpuset.

Forsøk 3.1 parameter: n-gram =2 ord, MFW=0/1000/100, *culling=0/100/20*, *list cutoff* = 3000, *consensus tree*=0.5 og *classic delta*. Se graf D.1 i D (s. 99).

Forsøk 3.2 parameter: n-gram =2 ord, MFW=0/0/0, *culling=0/100/20*, *list cutoff* = 3000, *consensus tree*=0.5 og *classic delta*. Se graf D.2 i D (s. 99).

Ved å sammenligne dette med et tilsvarende søk i “Likelangtkorpuset” kan man få svar på kutting av tekstlengden påvirker utfallene i undersøkelser sammenlignet med “Novellekorpus”. Resultatene blir beskrevet i 4.4 (s. 40).

Grafene i D.1 og D.2 på s. 99 og s. 100 er påfallende forskjellige. Forsøket med "Novellekorpuset" er mer presist gruppert enn i forsøket i "Likelangtkorpuset". Forsøket i "Novellekorpuset" er riktig gruppert i alle tilfellene, men grenene med hver forfatter er i noen tilfeller forgrenet sammen med en annen forfatter. Blant annet er Hagan og Hamilton forgrenet sammen, noe som også forekommer i C.1 (s. 91) som undersøkte bokstav-grammer (n=3). Dette kan indikere at det finnes fellestrekk mellom Hagan og Hamilton som modellen identifiserer.

En likhet mellom grafene D.1 (s. 99) og D.2 (s. 100) er at Klippenvåg er gruppert likt. To av tekstene er forgrenet sammen og den ene teksten er en utstikker under de andre på samme gren. Dette understøttes av andre grafer med parametrene: ord-gram (n=2) og MFW=0/1000.

Grafen D.2 på s. 100 fra "Likelangtkorpuset" er påfallende lik grafen C.2 i C på s. 92. Dette er ikke uventet i og med at forsøkene har like parameter med MFW = 0/1000/100, *bootstrap consensus tree*=0.5 og bruker samme korpus. I begge delene er Enger mest feilgruppert. I den ene er Enger plassert delvis sammen med Lindell. Johnsgaard er ikke gruppert på samme gren, men ved siden av hverandre på hver sin gren. I begge grafene er den ene Klippenvåg delvis riktig gruppert. Delvis riktig gruppert vil si at den ene teksten er på samme hovedgren som de andre to tekstene, men plassert på en egen, lavere plassert gren. Den eneste forskjellen mellom figurene er at tekstene til Ragde er gruppert ved siden av hverandre i D.2 og sammen i C.2. Antakeligvis er dette fordi forskjellen er at det i figur D.2 ble *culling*=0/100/20 brukt, i motsetning til i figur C.2.

4.4.2 Forsøk med *sampling*

Figur D.3 (s. 101) representerer hver prikk en del av tekstene med forskjellige farger som representerer hver forfatter. På grunn av antallet prikker er ikke forfatternavnene lagt inn i grafen, men farget. Kodene til forfatterne ligger i en tabell ved siden av graf D.3 (s. 101). Grafen illustrerer at enkelte forfattere er mer samlet enn andre: Oterholm (mørkelilla), Hagan (grønn), Hamilton (blå), Johnsgaard (svart), Lindvåg (mørkerød) og Lindell (grå). Enkelte andre er mer spredt: Klippenvåg (lilla), Enger (rød) og Ragde (lyseblå).

Kiøsterud (oransje) har færrest prikker i grafen, men de kan sees som delvis samlet. Denne spredningen er ut til å understøtte funn fra tidligere undersøkelser; Enger, Klippenvåg og Ragde er de forfatterne som oftest blir feilgruppert (jfr. figurene C.1, C.2, C.5, C.7 og D.2). I de fleste figurene er bare Enger og Klippenvåg feilgruppert. Unntaket er D.2 hvor alle tre feilgruppert. Et fellestrekk ved alle grafene er at de hadde MFW=0/1000/100 eller MFW=1000/1000. En sammenligning mellom *sampling*-figurene D.3, D.4 og D.5, viser en lignende avstand mellom Enger i alle tre figurene. Et annet sammenfall er at Oterholm ligger langt fra de andre forfatterne i alle tre forsøkene. Dette, i tillegg til sammenfallene mellom de andre figurene, indikerer at *random sampling* og *no sampling* gir tilnærmet like resultater, i alle fall når *list cutoff* er satt til 4000 ord. *Normal sampling* er interessant fordi den forteller om alle delene av teksten ligger nært hverandre visualisert i en graf. Spesielt spennende er Enger sine tekster i figur D.3 fordi tekstene viser hvorfor han ofte kan bli feilgruppert. Likheter og ulikheter i grafene viser frekvensene til de 1000 mest frekvente ordene/n-grammene og hvor langt enhetene står fra hverandre i henhold til frekvensene.

4.5 Forsøk med "Kjønnskorpuset"

"Kjønnskorpuset" består av 21 tekster á 4500 ord av 10 forskjellige forfattere, 5 menn og 5 kvinner. Målet med å lage korpuset er å forsøke en kjønnsbasert forfatterprofilering.

Forfatter	Kjønn
Ragde	K
Johnsgaard	M
Hagan	K
Lindell	K
Oterholm	K
Klippenvåg	M
Lindvåg	K
Enger	M
Hamilton	M
Kiøsterud	M

Tabell 4.4: Oversikt over forfatternes kjønn

Tekstene ble sendt gjennom programmet *Stylo* både med statistiske metoder og maskinlæring. Trekkene som ble valgt var ord- og bokstavgram.

For å utføre overvåket læring på korpuset i *Stylo* måtte korpuset deles inn i to mapper. En med treningssettet og en med testsettet. Mappene fikk navnene *primary_set* og *secondary_set*.

Tabell 4.5: Oversikt over forsøk i Kjønnskorpus

Forsøk nr.	Algoritme	Distanse	n-gram	MFW	Culling
5.1	Cluster Analysis	Classic Delta	3-bokstavgram	100/100/100	0/0/0
5.2	Cluster Analysis	Classic Delta	3-bokstavgram	0/0/0	0/0/0
5.3	SVM	Classic Delta	3-bokstavgram	100/100/100	0/0/0
5.4	SVM	Classic Delta	3-bokstavgram	0/0/0	0/0/0
6.1	Cluster Analysis	Classic Delta	2-ordgram	100/100/100	0/0/0
6.2	SVM	Classic Delta	2-ordgram	100/100/0	0/0/0

4.5.1 Bokstavgram i "Kjønnskorpuset"

Det første forsøket gikk ut på å bruke statistiske metoder for å gruppere tekstene. I dette forsøket valgte jeg ut trigrammer. Statistikkmessig valgte jeg *cluster analysis* som produserer et dendrogram. *Cluster analysis* er et godt valg dersom man ønsker en enkel analyse og ikke flere itereringer (Eder et al., 2013). Krav til *cluster analysis* er at MFW innstillingene min. og maks. er like. Det samme gjelder innstillingene for *culling* min. og maks.

Andre parametre: MFW=100/100/100, *culling*=0/0/0, *list cutoff* = 10 000, *no sampling*. Deretter ble en annen test kjørt hvor MFW=0/0/0 (jfr. tabell 4.5 for en oversikt).

I maskinlæringsforsøket ble det utført en lik test med det samme parametrene, untatt når det kom til de statistiske metodene. SVM med *Classic Delta* ble anvendt i forsøkene. SVM brukte lineær kernel, anbefalt av Eder et al. (2013) fordi antall variabler er betydelig større enn klassene.

Resultatet ble 90 % riktig med SVM.

Deretter ble en annen test, med de samme parametrene utført med MRW=0/0/0. Resultatene her var 50 % riktig klassifisering med SVM.

4.5.2 Ordgram i "Kjønnskorpus"

I dette forsøket ble de samme parametrene som i n-gram-forsøket anvendt, med MFW = 100/100. Den eneste forskjellen var at i stedet for å undersøke bokstav-gram (n=3) ble ord-gram (n=2) undersøkt (jfr. tabell 4.5 med 6.1 og 6.2).

Mangelen på kutting av listen n-grammer var tydelig i forsøket med ikke-overvåket læring, det tok svært lang tid å fullføre forsøket i *Stylo*.

Det overvåkede læringsforsøket hadde de samme parametrene som det forrige forsøket, men brukte SVM, med lineær kernel.

Resultatet var 90 % korrekt klassifisering, med 9/10 korrekte klassifiseringer. Den eneste feilklassifiseringen var Unni Lindell sin tekst "Slangebæreren".

4.5.3 Funn i undersøkelsene i "Kjønnskorpuset"

I forsøket på s. 42 ble forskjellige trekk i "Kjønnskorpuset" undersøkt for å se om korpuset og leksikalske trekk egnet til å gruppere forfatterkjønn. Ikke-overvåket og overvåket læring ble kontrastert opp mot hverandre. Til sist ble MFW eksperimentert med.

Bokstav-gram

I 4.5.1 (s. 42) ble trigrammer undersøkt. Undersøkelsen viste en betydelig forskjell mellom resultatene, når MFW ble først satt på 0 og deretter på 100. Med *cluster analysis* varierte dendrogrammene betydelig. Det første diagrammet er innstilt på 100 MFW og det andre på 0, se grafene i E (jfr. s. 105).

Forfattermessig ser 100 MFW ut til å være verdien som gir best resultater, sammenlignet med 0 MFW.

Ved å bruke SVM er det en tydelige forskjell mellom 100 MFW og 0 MFW. MFW=100 gir SVM en 90 % riktig klassifisering, i motsetning til MFW=0 som ga en 50 % riktig klassifisering.

Tabell 4.6: Feilklassifisering i SVM (0 MFW)

Klasse og tekst	Feilklassifisering
K_Lindell_Slangebæreren	M
K_Ragde_EnKald	M
M_Enger_Solformørkelse	K
M_Hamilton_Tiggerkongen	K
M_Klippenvåg_Body	K

Ved å sammenligne feilklassifiseringen av forfatternes kjønn mellom de ulike metodene er det mulig å se noen likheter. En sammenligning av klassifiseringene med figuren E.2 (s. 106) gjør

det tydelige at Enger, Hamilton og Klippenvåg er tydelig feilklassifisert i begge forsøkene. Med *cluster analysis* er alle tre plassert nærmest det motsatte kjønn. Ragde sin tekst står for seg selv, men ligger også nærmest tekstene til mannlige forfattere. En forskjell fra forsøkene med SVM og *cluster analysis* er at Lindell er feilklassifisert i SVM-forsøket, men ikke med *cluster analysis*. Med *cluster analysis* er teksten hennes plassert nærmest sin egen tekst. I tillegg er Oterholm og Lindvåg mer feilgruppert med *cluster analysis* enn i SVM.

Ordgram i kjønnskorpus

Resultatene i bigram-forsøket av ord var i stor grad lik som i undersøkelsen med bokstav-gram. Ut fra *cluster analysis* er det vanskelig å si om grupperingen etter kjønn ga positivt resultat eller ikke. SVM ga 90 % riktig klassifisering, med "K_Lindell_Slangebæreren" som den eneste feilklassifiseringen. Med *cluster analysis* hadde heller ikke direkte feilgruppert Lindell, akkurat som i bokstav-gram forsøket, se graf i E.1 (s. 105) og figur E.3 (s. 107).

4.6 Dokumentasjon av forsøkene og resultatene

På grunn av størrelsen til dokumentasjonen av forsøkene er deler ekskludert fra tekst og vedlegg. I stedet er den gjort tilgjengelig via internett. Dokumentasjonen som er lastet opp er *Stylo*-genererte konfigurasjoner til forsøkene, grafer og ord- og frekvenslister. Den er tilgjengelig via linken:
`<figshare.com/s/446652b6f8c211e4bc2106ec4b8d1f61>`.

4.7 Diskusjon og konklusjon

Forsøkene i studien hadde et bredt fokus. Fokuset var på å undersøke hvilke metoder som hadde en effekt på forfatterattribuering. Resultatene ble ofte illustrert i grafer, når statistiske analysemetoder ble anvendt. Disse metodene uttrykte ofte underliggende strukturer og klynger av tekstene. Av den grunn blir tolkningene av grafene subjektive.

Diskusjon

N-grammer viste seg til jevnt over å være effektive til å forutsi forfatterskap, men ikke kjønn. Enkelte parametre og sammensetninger av parametre påvirket resultatene drastisk. Eksempelvis var MFW interessant fordi endret resultatene i betydelig grad ved manipulering av ordlisten.

Enkelte parametre og valgmuligheter ble i denne studien ikke anvendt, blant annet: liten/stor bokstavskille og sletting av pronomen. Grunnen var at studien måtte avgrenses for å ikke bli uoversiktlige og det ble på forhånd antatt at andre parametre ville være mer effektive å teste ut. Det å teste ut alle valgmulighetene på norske pronomen ville økt størrelsen på forsøkene i større grad enn hva som var mulig i denne studien. Derfor ble ikke n-gram verdiene endret mer enn mellom 2 og 3, på bokstavgram og ordgram. n=1 for ord ble unngått for å delvis kunne utelukke tematisk påvirkning av høyfrekvente ord. Det mest effektive valget av n er potensielt interessant studie, som kan

gi innsikt i hvilken informasjon de statistiske metodene finner nyttig. Eksempelvis er n=1 for ord effektivt, delvis fordi trekket inneholder tekstkarakteristiske innholdsord.

Som tidligere nevnt påvirket MFW-parameteret resultatene betydelig. MFW hentet ut fra tekstene de mest frekvente ordene (eller n-grammene) og sammenlignet de med mest frekvente ordene (eller n-grammene) fra de andre tekstene. Forsøk med et MFW parameter indikerer mer korrekte grupperinger, enn forsøk uten MFW. Grunnen til forbedret gruppering er trolig at frekvensen til de mest frekvente ordene (eller n-grammene) er mer karakteristiske for en tekst enn mindre frekvente ord (eller n-grammer). Dersom man antar at mindre frekvente ord er mindre karakteristiske vil en studie med utelatelse av denne typen ord kunne forvente enn forbedring i forhold til testing med de mindre frekvente ordene. MFW-frekvensene er ikke kun karakteristisk til en tekst, men er karakteristisk til forfatteren bak teksten.

I forsøkene ble grensen for MFW satt til 1000 ord eller en økning til 1000. Det ville vært interessant å undersøke et mindre sett med MFW. I "Kjønnskorpus"-forsøkene var MFW=100 og ut fra en forfatterattribueringsvinkel presterte tilsvarende eller bedre resultater enn med forsøk hvor MFW var høyere.

I 4.3.3 med forsøkene med *culling* og MFW ble ofte Oterholm og Lindell gruppert riktig og Enger feil. Dersom man ser bort fra mulige påvirkninger av sjanger- og tema, åpner dette opp for at forfattere kan sees på en skala, som skiller lett gjenkjennelige forfattere fra mindre gjenkjennelige forfattere ut i fra ord-gram. Skalaen kunne indikert hvor særegne frekvensene til en forfatter var i forhold til andre forfattere. Den ville også kunne indikert konsistente og karakteristiske ordvaner, siden ordgram (n=2) blir brukt.

I *culling*-forsøkene i 4.3.3 påvirket ikke *culling* resultatene i tydelig grad. I forsøkene ble parameteret *culling*=0-100 brukt. Det kan tenkes at andre forsøk hvor parametrene er satt til andre verdier, eksempelvis *culling*=20, vil være mer effektivt. Til nå har ikke *culling* med økning fra 0/100/20 vist seg å være synlig effektivt i forsøkene som er gjort. Det *culling* forsikrer om er at ord som er karakteristiske for en tekst blir utelatt av analysen og dermed minsker en eventuell temapåvirkning.

I forsøkene i 4.3.3 ble Enger ofte feilgruppert med *bootstrap consensus tree*. Det kan tenkes at ord-gram og bokstav-gram ikke presterer å identifisere eller gruppere skrivestilen til Enger. En medvirkende faktor kan være at tekstene hans står langt fra hverandre stilmessig og dette virker inn på resultatet. *Sampling*-forsøkene kan belyse dette. I *sampling* forsøkene i 4.4.2 (jfr. figur D.3, s. 101) med *normal sampling* kan indikere hvorfor Enger ofte blir feilgruppert (jfr. f. eks. figurene D.2, C.6, C.5, C.4, C.2, C.3,C.1). Enger ble i figur D.3 inndelt i mindre tekstbiter på 4000 ord. I figuren er prikkene som representerer Enger mer spredt i forhold til de andre forfattere. Et nærmere blikk på figuren kan indikere to svake Enger-*clusters*, som kan representere Enger sine 2 tekster. Klyngene åpner for at tekstene, sjangermessig eller tematisk står langt fra hverandre i innhold. En annen mulighet er at Enger har et variert ordbruk innad i tekstene. En separat analyse av Enger må til for å kunne gi svar på dette. Det kan konkluderes med at inndeling av tekstene i mindre deler kan indikere hvorfor enkelte forfattere oftere blir feilgruppert enn andre.

Det var tilstrekkelig sammenfall mellom *sampling*-forsøkene, *normal sampling* og *random samling* til å gi tilnærmet like resultat. Trolig betyr sammenfallet at frekvensenene av n-gram i tekstene

er gjennomsnittlig like. En lavere *sampling*-ordgrense kunne gitt andre frekvenser og variasjon innad i *sampling*-forsøkene, fordi frekvensene blir mindre representative for hele teksten. Eksempelvis indikerer Enger at n-gram frekvensene kan variere mye innad eller mellom tekster.

I 4.4.1 på s. 40 ble "Novellekorpuset" og "Likelangtkorpuset" sammenlignet for å undersøke om tekstlengde påvirket resultatene. "Novellekorpuset", som ikke hadde kutt i tekstlengden presterte å gruppere forfatterne sammen med sine egne tekster. Derimot ble Enger gruppert feil i forsøket med "Likelangtkorpuset". Dette indikerer at kortere tekstlengde kan minimere presisjonen til de statistiske modellene, selv når høyfrekvente trekk blir anvendt.

I "Kjønnskorpus"-forsøkene i 4.5 ble både statistiske metoder og maskinlæring brukt. Figurene E.1, E.2 og E.3 visualiserer en *cluster analysis*. I ingen av figurene er det tydelig at modellene har gruppert forfatterne etter kjønn. I forsøket med SVM ble forfatterne klassifisert 90 % riktig med MFW=100 og 50 % riktig med MFW=0. Dette tyder på at MFW kan ha stor innvirkning på resultatene når man bruker SVM. En nærmere undersøkelse med forfatterattribuering ville gitt en bedre indikasjon på påvirkningen av MFW, fordi det er usikkert om forfatterne er gruppert ut i fra forfatter eller kjønn. Sammenlignet med forsøkene med *cluster analysis* er grupperingen tilsynelatende gjort ut fra forfattermessig likhet. Resultatene av forsøket viser at tekstene trolig er gruppert ut fra en *forfatteravhengighet* og ikke *kjønnsavhengighet*. Med andre ord representerte frekvensene ikke kjønn men forfatter og derfor ble forfatterne gruppert på den måten de ble gruppert i "Kjønnskorpus"-forsøkene. Dette var ikke uventet. Korpuset var laget med hensikt til forfatterattribuering. Antageligvis var det ikke et tilstrekkelig antall tekster eller forfattere til å kunne generalisere frekvensene utover et forfatternivå.

I forsøkene ble det kun brukt leksikalske trekk, en programbegrensning. Bare n-gram ble brukt, selv om ordlister også kunne brukes. Eksempelvis åpnet programmet for at funksjonsord kunne brukes, men jeg valgte å fokusere på andre parametre i denne studien.

Stylo kan være et velfungerende og brukervennlig verktøy for stilometriske undersøkelser. Pakken krever lite datakunnskaper og preprosessering. Til undersøkelsene her laget jeg programmet "wordsplitter.sh", som et eneste supplementet for å korte ned på tekstene.

Det hadde vært ønskelig med muligheten for kryssvalidering ved valg av maskinlæringsalgoritmer. Kryssvalidering ville minske potensiell skjevhet og variasjon i resultatene ved å trene og teste på hele datasettet. Med *Stylo* ble maskinlæringsalgoritmen trent på 11 tekster testet på 10 tekster og tekstene ble ikke gjensidig trent og testet på hverandre.

En videre utvikling av programmet til å inkludere norsk i pakken for å kunne fjerne pronomener i teksten ville vært interessant for å undersøke om det kan påvirke resultatene i forfatterattribueringsundersøkelser.

Konklusjon

Undersøkelsene i dette kapittelet hadde betydelig variasjon. Undersøkelsene varierte i korpus, problemstilling (forfatter og kjønn), parametre og statistiske metoder. Undersøkelsene kunne med fordel hatt et smalere fokus. For eksempel kunne fokuset vært kun på forfatterattribuering og utelatt

"Kjønnskorpuset". Et smalere fokus ville gjort det mulig å undersøke i hvilken grad forskjellige størrelser av MFW-parameteret påvirker resultatene, i hvilken grad tekstlengde påvirket forsøkene eller hvilke n som er mest informativt. Til gjengjeld har det brede fokuset gitt en oversikt over effekten til enkelte parametre, hvorfor noen forfattere ofte ble gruppert eller klassifisert feil og hvor lite egnet "Kjønnskorpuset" er til forfatterprofilering.

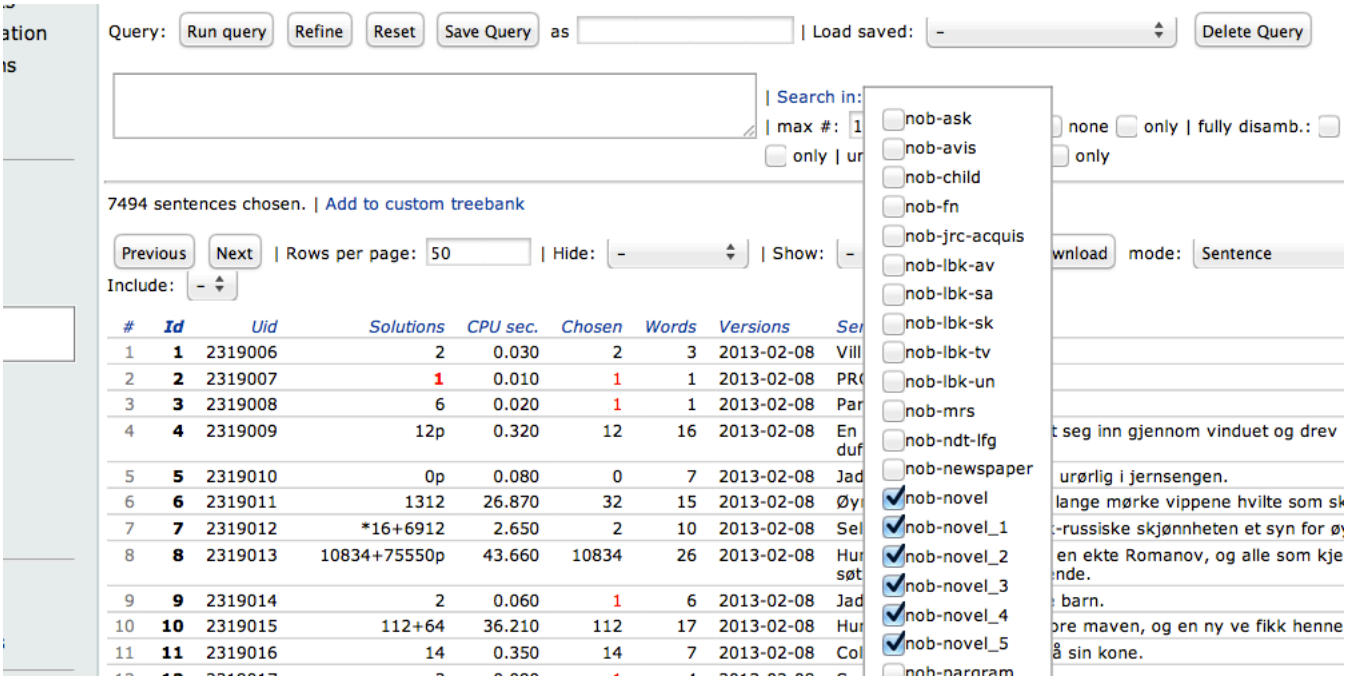
Kapittel 5

INESS - trekk og forsøk

INESS-forsøket består av tre hoveddeler: Korpus, preprosessering og statistisk modellering. Tekstene fra korpuset mitt lå i INESS-trebanken og der de ønskede frekvensene ble hentet ut herfra. Preprosesseringen besto av søk i trebanken, behandlingen og klargjøringen av frekvensene. Til sist modellerte maksinlæringsprogrammet WEKA med forskjellige statistiske modeller frekvensene.

5.1 Korpus til INESS-forsøk

Tekstene ble hentet fra de seks "nob_novel_0-5" i INESS. Ved å hake av disse korpusene og å søke på titlene til tekstene og søkeuttrykk kom trekkfrekvensene hentes ut og føres over i tabeller, om vist i figur 5.1.



Figur 5.1: Avkrysning av aktuelle INESS-korpus

Eksempel på søk som ble brukt for å hente alle tekstene ut:

```
#x >(PASSIVE) "\+" :: title = "(En kald.*|Ansikt.*|Vansk.*|Veiv.*|.*vakker|
Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|Bruck.*|Et vir.*|
Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|Kjærligheten.* triumf|Tigge.*|
Ved fj.*|Den n.*)"
```

Figuren 5.2 viser et eksempel på søket for passiv. I den første kolonnen, *Count* vises frekvensen av passiv for hver av tekstene. I neste kolonne, *title*, er navnet på tekstene.

Figur 5.2: Eksempel på søk og resultat i INESS

5.2 Trekkutvalg

I de foregående forsøkene med programmet *Stylo* ble et begrenset antall leksikalske trekk brukt. I denne delen av undersøkelsen brukes andre typer trekk. Jeg ønsker å ekstrahere ut syntaktiske trekk fra INESS, for å utvikle et enkelt sett med trekk. Trekkene ønsker jeg deretter legge i en vektor og kjøre gjennom en datamaskinell læringsmodell. I 3.1.2 beskrev jeg kort INESS-trebanken og krav til tekster som skal benyttes til stilometriske formål. Det mest sentrale er valget av trekk som skal undersøkes og valg av søkeuttrykk for å oppnå dette. Deretter vil jeg diskutere kravene som må stilles til frekvensene man får ved søk i INESS.

Trekkene fra INESS er interessante å undersøke fordi INESS-trebanken gir mulighet til å hente ut trekk som ikke tidligere er blitt benyttet i tidligere forskning.

Siden trekkene kan være særnorske, kan de indirekte bli påvirket av subnormer i bokmål. Et eksempel på en valgfri norm kan være fraværet av hunkjønn hos enkelte bokmålsbrukere som kan være med på å karakterisere en skrivestil.

Trekkvalget er eksperimentelt og er motivert ut ifra et ønske om på å finne trekk som indikerer forfatterskap.

I tillegg er det viktig at trekkene er høyfrekvente nok til å kunne brukes. Hadde de vært for lavfrekvente kunne tilfeldigheter ha påvirket frekvensene i altfor stor grad. For å unngå dette problemet ble tilfeller hvor flere av tekstene hadde null forekomster i et søk utelatt. Et eksempel er akkustativ, som ble utelatt til tross for annotering for kasuset i INESS.

Det er til tider vanskelig å skille mellom trekk som er diskursavhengige eller semantisk avhengige. Det kan også være vanskelig å avgjøre i hvor stor grad et trekk er innholdsavhengig. Med innholdsavhengig menes egenskaper i teksten som er avhengige av tekstens sjanger eller tema.

Trekkvalget er delvis preget av hensiktsmessighet. Enkelte begrep i setningsanalysene i trebanken er ikke selvforklarende og ble dermed valgt bort. For eksempel ble adverb valgt bort fordi flere av adverb-typene som var annotert ikke var dokumentert godt nok eller intuitivt forståelige for en utenforstående.

5.2.1 Frekvenser i INESS

Den første utfordringen med søk i INESS er at ikke alle setningene i trebanken er parset og disambiguerte. Dette gjør det vanskelig å bruke frekvensene fra søkene uten å vite representativiteten til søket. Dersom søket på antall hankjønn substantiv, uten å vite om alle setningene i teksten er fullstendig disambiguert, kan man ikke anta at frekvensen av søket samsvarer med reell frekvens. Derfor må søkefrekvensene ikke ansees som absolutte, men som relative.

Denne relativiteten må også reflektere representasjonene i teksten i størst grad mulig. Derfor vil en frekvens av hankjønn substantiv måtte settes opp mot en annen tilsvarende relativ frekvens. Det vil si at det i stedet for å sette substantiv hankjønn opp mot totalt antall substantiv i teksten, ville det være gunstigere å sette det opp mot totalt antall substantiv med registrert kjønn per tekst.

Dette forutsetter at en relativisering er representativ i samme grad på tvers av alle tekstene. Det antas at alle tekstene er disambiguerte og parset på tilnærmet samme måte og i samme grad. Dette er grunnen til at bare tekster med over 80 % parsing er brukt i disse forsøkene. Dette forminsker et potensielt representativitetsproblem.

Forslag til relative frekvenser som kan brukes er:

- **Binære søk:** Dette er søk hvor to motsetninger settes opp mot hverandre for å måle den relative frekvensen. Et eksempel på dette er forholdet mellom entall og flertall. Et binært forhold forutsetter at det bare finnes to mulige utfall av søket. Det ville for eksempel uriktig å bare måle forholdet mellom hunkjønn og hankjønn, uten å inkludere intetkjønn.

- **Antall forekomster/antall forekomster i samme kategori:** Gitt en kategori har flere enn to trekk kan det være nyttig å relativisere frekvensen i forhold til forekomster av kategorien

totalt. Eksempelvis hvis hunkjønn-frekvensen skal relativiseres kan dette gjøres ved: hunkjønn/(hunnkjønn+hankjønn+intetkjønn).

- **Antall forekomster/Antall mulige forekomster:** Et eksempel på denne typen forhold kunne vært antall adjektiv målt opp mot antall *mulige* adjektiv i en tekst. Med mulige adjektiv menes det potensielle steder i en tekst der det kan stå et adjektiv. Hovedutfordringen med denne typen forhold er at antall mulige forekomster må kunne være finitte og målbare. I norsk kan flere adjektiver stilles etter hverandre i en setning. Dette blir dermed ikke en ideell frekvens, fordi mulige forekomster ikke lar seg måle.

- **Antall forekomster/Antall setninger, leddsetninger eller ord:** Dette er en mer absolutt type søk. Her blir ikke to eller flere søk av samme type satt opp mot hverandre. For eksempel kan det være interessant å se på antall leddsetninger i forhold til antall setninger totalt. Det forutsetter at antall leddsetninger er målbart og parset i høy nok grad til å gi en representativ frekvens når det er satt opp mot en absolutt frekvens.

For å konkludere er valget av relativ frekvens for å representere forekomstene i INESS utfordrende. Det må tas høyde for hvor mange mulige motsetninger som finnes til det søket man gjør og om det er mulig å måle fraværet av forekomster og antall mulige forekomster. For å få best mulig representativitet er det også viktig å sikre at de frekvensene man setter opp mot hverandre er disambiguert i omtrent samme grad.

5.2.2 Trekk og søk

Grammatisk kategori: Bestemthet

Bestemthet baserer seg på om noe/noen er referert til tidligere i teksten eller ikke, eller at leseren på forhånd har kjennskap til referenten og undersøkes fordi det er en grammatisk kategori som kan antas å være diskursavhengig til en viss grad. På en annen side kan det tenkes at bestemthet brukes som et virkemiddel av forfatteren til å gi en illusjon av kjennskap.

I søkene mellom bestemt og ubestemt valgte jeg å benytte en binær frekvens for å relativisere søkene.

Søk i INESS:

```
[frame=single] #x >DEF "\-" & #x >(NTYPE NSYN) 'common' :: title = "
(En kald.*|Ansikt.*|Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|
Avbr.*|Avslut.*|Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|
Hvis noen s.*|Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

I dette søket ble kun substantiv av typen 'common' i INESS benyttet. 'common' er kontrastert med trekkene 'pronoun' og 'proper', men i sammenheng med søk på bestemthet av substantiv var ikke disse interessante eller nødvendige å inkludere. Søket fanger heller ikke opp bestemthet i adjektiv, fordi adjektivenes bestemthet kun er relevant i setninger hvor sambøyningsubstantivet til et adjektiv er utelatt, eksempelvis: "Den gule der borte".

Grammatisk kategori: Tall

Antageligvis er frekvensen av entall eller flertall dels eller helt avhengig av innholdet i teksten. Det er ikke noe som indikerer at forfattere har en preferanse for å skrive i eller om enten entall eller flertall. Dermed er tall mest sannsynlig innholdsavhengig. Jeg velger likevel å undersøke tall likevel fordi jeg ønsker å bekrefte hypotesen om at tall ikke er et nyttig trekk.

Tall har to mulige uttrykk, og det er derfor mulig å sette dem i et binært forhold og finne en relativ frekvens. Søket skiller ikke mellom ordklasser.

Søk i INESS:

```
#x >(NUM) 'sg' :: title ="(En kald.*|Ansikt.*
|Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|
Avslut.*|Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|
Hvis noen s.*|Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Grammatisk kategori: Kjønn

Kjønn kan være et interessant trekk fordi det kan brukes enten som et trekjønnssystem eller tokjønnssystem i bokmål (Faarlund et al., 1997, s. 150). I tokjønnssystemet går hankjønn og hunkjønn sammen til et felleskjønn. Systemet bygger på dansk og brukes i riksmål og kan brukes i konservativ bokmål (Faarlund et al., 1997, s. 151). I tillegg brukes tokjønnssystemet i bergensk tale, noe som trolig reflekteres i skriftspråket til bergensere.

Frekvensmessig må det tas høyde for alle tre motpartene. Derfor er det hensiktsmessig å lage en relativ frekvens med for eksempel: hunkjønn/(hankjønn+hunkjønn+intetkjønn).

Foruten "FEM", er de andre kategoriene i søket: "MASC" og "NEUT". Det blir ikke her søkt etter fraværet av trekkene, fordi frekvensen baserer seg på de eksisterende tilfellene av trekkene. Et kjapt søk indikerer at det finnes flere NEUT- enn det finnes tilfeller av FEM+ og MASC+ til sammen. For eksempel har teksten "En kald dag i helvete" 8010 av FEM+ og MASC+, men 11393 tilfeller av NEUT-. Det er uklart hvorfor frekvensene er inkonsekvente, men dette er grunnen til å relativisere frekvensen på grunnlag av de eksisterende tilfellene av trekkene og se bort ifra søk av typen NEUT-.

Søk i INESS:

```
#x >(FEM) "\+" :: title = "(En kald.*|
Ansikt.*|Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*
|En gruso.*|Avbr.*|Avslut.*|Body.*|Bruck.*|Et vir.*|
Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|Kjærligheten.* triumf|
Tigge.*|Ved fj.*|Den n.*)"
```

Grammatisk kategori: Diatese

INESS er annotert for passiv. Det er usikkert om passiv er semantisk motivert eller motivert ut ifra et annet grunnlag, for eksempel en diskurs- eller fokuseringspreferanse hos forfatter. Derfor velger jeg å utføre en undersøkelse av dette trekket med forbehold om uvisshet av motivet for bruken av passiv.

En svakhet med passiv er at INESS ofte ikke annoterer når setninger ikke er passive. Eksempelvis består Ragdes "En kald dag i helvete" av 5560 setninger, 10 er annotert for passiv og 428 er annotert for ikke-passiv. Dette kan være problematisk når man ønsker et representativt forhold mellom disse to frekvensene. Det er mulig å anta at passiv i like stor grad er annotert for i alle tekstene og ikke-passiv i like liten grad er annotert i tekstene. Dermed blir forholdet sammenlignbart på tvers av forfatterne. Med dette forbeholdet kan man lage et lage en relative frekvensen mellom passiv og ikke-passiv.

Søket av passiv avslørte også hva som skjer dersom et trekk ikke er annotert i en av tekstene eller ikke forekommer; teksten dukker ikke opp i søket. Til tross for at søket var lavfrekvent i noen tekster, og helt fraværende en, valgte jeg å beholde det.

Søk i INESS:

```
#x >(PASSIVE) "\+" :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|
Avslut.*|Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Referanse

Referanse vil i dette tilfellet si om det annoterte ordet referer til noe som er nevnt tidligere i teksten eller ikke. Referanse er annotert for trekkene, '+' og '-'.

Det er interessant å undersøke forekomstene av referanser i en tekst siden dette er avhengig av diskursen i teksten og viser til tidligere eller kjente objekter. På denne måten minner referanse bestemthet.

For å danne en relativ frekvens kan referanse/ikke-referanse settes i et binært forhold.

Søk i INESS:

```
#x >REF "\+" :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|
Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Adjektivtype: attributativ/predikativ

INESS er annotert for to forskjellige adjektivtyper: "attributativ" og "predikativ". Disse er kan være de samme adjektivene, men har ulike syntaktiske plasseringer.

Eksempel:

A ”Ei lang bok”. (Attributativ)

B ”Boka er lang”. (Predikativ)

De har et binært forhold som kan være et godt trekk for å indikere forfatterskap, siden de både er tilsynelatende valgfrie, gjensidig utelukkende og innholdsfrie. Søket er relativisert ved å sette forholdene opp binært.

Søk i INESS:

```
#x >(ATYPE) 'predicative' :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|
Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Koordinering: Totalt, og, eller, men og komma

Frekvensen av koordineringsformer kan være interessante trekk fordi de ikke er innholdsavhengige, men avhengige av syntaksen i en setning. Formene jeg har valgt å søke etter er høyfrekvente og ikke obligatoriske i en setning. Lavfrevente former ble utelukket.

Dersom koordinerende former viser seg å være gode trekk, kan det være spennende å inkludere mindre frekvente former på et senere tidspunkt. Totalt antall koordineringsformer ble tatt med som et eget trekk, i tillegg til de fire andre.

For å relativisere frekvensene ble søkene satt opp mot alle forekomstene av koordinering til sammen. Totalt antall koordineringsformer ble målt mot antall ord i en tekst for å relativisere frekvensene.

Søk i INESS:

```
#x >(COORD-FORM) 'og' :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*
|Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Kasus: Nominativ, oblik og genitiv

Kasus er delvis semantisk og syntaktisk avhengig. Genitiv indikerer eierskap og er dermed avhengig av innholdet i en tekst. Nominativ og oblik indikerer subjekt og direkte objekt i en setning. Disse er delvis avhengige av predikatet og dermed indirekte påvirket av det innholdsmessige. Oblik er ikke nødvendigvis obligatorisk i en setning og kan dermed være et interessant trekk å måle frekvensen til. For eksempel kan en lav andel nominativ bidra til å indikere en høy andel passive setninger, på grunn av mangelen på subjekt.

Kategorien kasus har 5 forskjellige trekk annotert i korpusene jeg søkte i: Oblike, nominativ, akkustativ, dativ og genitiv. For å relativisere frekvensene ble trekkene kontrastert med de resterende trekkene. Akkustativ var kun annotert 7 ganger totalt i 2 av 21 tekster og ble derfor utelatt. Dativ var ikke annotert for i de tekstene jeg brukte og ble utelatt. Dermed valgte jeg å benytte kun genitiv, nominativ og oblik.

```
#x >(CASE) 'obl' :: title = "(En kald.*|Ansikt.*|Vansk.*|
Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Liste over relative frekvenser fra søk

Frekvensene generert av INESS er lagt som tabeller i vedlegg I (s. 117). Tabellene viser forekomstene av søkene før de er omgjort til relative frekvenser.

5.3 Preprosessering - klargjøring til modellering

Frekvensene fra alle søkene ble satt i en tabell, i en .csv-fil. Der i fra ble tabellen hentet inn i R:

```
> setwd("/Users/victoriatroland/Desktop/INESS korpus/INESSforsøk/"")
> megatable = read.csv(file="INESSFrekvenser.csv", header=TRUE, sep= "\t")
```

Fjernet frekvensene som ikke skulle være med:

```
> megatable <- megatable[c(1,2,3,4,8,9,10,11,12,15,16,17,18,23,24, 25,26,
27,28,29, 30, 31, 32, 33, 57, 58, 59, 60, 62,63)]
```

Skrev dette over i en ny fil:

```
> write.csv(megatable, file="cutfile.csv")
```

Hentet inn den nye filen, som er komma-separert.

```
> table <- read.table(file="cutfile.csv", header=TRUE,sep=",")
```

Laget relative frekvenser av trekkene:

```
> forf <- table\$Forfatter
> coordtot <- (table\$COORD.FORM.tot/table\$Ord)
> coordmen <- (table\$COORD.FORM.men/table\$COORD.FORM.tot)
> coordkomma <- (table\$COORD.FORM.comma/table\$COORD.FORM.tot)
```

```
> coordeller <- (table\$COORD.FORM.eller/table\$COORD.FORM.tot)
> coordog <- (table\$COORD.FORM.og/table\$COORD.FORM.tot)
> passiv2 <- (table\$Passive../table\$Passive...1)
> ref <- (table\$Referanse../table\$Referanse...1)
> atype <- (table\$ATYPE.predicative/table\$ATYPE.attributive)
> hunkj <- (table\$Fem../(table\$Fem..+table\$Neut..+table\$Masc..))
> intkj <- (table\$Neut../(table\$Fem..+table\$Neut..+table\$Masc..))
> hantkj <- (table\$Masc../(table\$Fem..+table\$Neut..+table\$Masc..))
> gen <- (table\$Case.gen/(table\$Case.gen+table\$Case.nom+table\$Case.obl))
> nom <- (table\$Case.nom/(table\$Case.gen+table\$Case.nom+table\$Case.obl))
> obl <- (table\$Case.obl/(table\$Case.gen+table\$Case.nom+table\$Case.obl))
> tall <- (table\$Pluralis/table\$Singularis)
> best <- (table\$Bestemt/table\$Ubestemt)

id <- c("ar1", "ar2", "mj1", "mj2", "ph1", "ph2", "ul1", "ul2", "ao1",
"ao2", "ok1", "ok2", "ok3", "el1", "el2", "re1", "re2", "dh1", "dh2",
"ek1", "ek2")
```

Sette disse nye variablene i samme tabell:

```
table2 <- data.frame(CoordTot = coordtot,CoordMen=coordmen,
CoordKomma=coordkomma,CoordEller=coordeller, CoordOg=coordog, Passiv=passiv,
Referanse=ref, AdjType=atype, Hunkjoenn=hunkj, Hankjoenn=hantkj,
Intetkjoenn=intkj, Genitiv=gen, Nominativ=nom, Oblik=obl, Tall=tall,
Bestemthet=best, ID = id, Forfatter=forf)
```

Skrev dette ut i egen fil:

```
> write.csv(table2, file="freqtableiness.csv")
```

Kuttet første kolonne:

```
cut -d, -f2-19 freqtableiness.csv > freq2.csv
```

Deretter ble headeren tatt ut og lagret i en egen fil:

```
head -1 freq2.csv > header.csv
tail -n +2 freq2.csv > freq3.csv
```

Punktum ble lagt til på slutten av hver linje:

```
awk '\{ print $0 "." \}' < freq3.csv > freq4.csv
```

Deretter ble filen delt i to manuelt. En linje som representerte hver forfatter gikk inn i en ny fil, testssettet, "testfreq3.csv". Resten gikk inn i treningssettet, "freqdivided.csv".

Randomisering av linjene

```
gshuf freq4.csv > frequencies2.csv
```

Limte inn igjen header

```
cat header.csv frequencies2.csv > wekafreq.csv
```

Eksempel av linje fra prosessert fil:

```
0.123269611074489,0.137112722478576,0.04614370468029,0.653263019116678,
0.00759493670886076,0.0616462762798432,0.609829488465396,0.0448751300728408,
0.494406867845994,0.460718002081165,0.0111644214799779,0.470843329027496,
0.517992249492526,0.29144137185231,2.5664585191793,ek3,Kiøsterud.
```

Hvert nummer representerer frekvensen til et trekk.

5.3.1 Omgjøring av frekvensene til diskrete verdier

Å gjøre om frekvensene til kvantiler er en måte å generalisere datasettet på. En omgjøring til kvantiler gjør datasettet om til diskrete verdier. Bakgrunn for komprimeringen av datasettet er å gjøre datasettet håndterbart for enkelte læringsalgoritmer. Algoritmer som kan ha positiv effekt av denne omgjøringen er *Naive Bayes*, k-NN og regelbaserte klassifikatorer, se kapittel 2. Nevrale nettverk og SVM presterer derimot generelt bedre med flere dimensjoner og kontinuerlige verdier (Kotsiantis, 2007).

Kvantilene er delt inn i kvartiler i forhold til resten av frekvensene i samme variabel. For eksempel er alle frekvensene til alle tekstene for "passiv" delt inn i 4 deler. De 1/4 laveste frekvensene får tallet 1, frekvensene mellom 1/4 og 2/4 får tallet 2, frekvensen mellom 2/4 og 3/4 får tallet 3 og frekvensen mellom 3/4 og 4/4 får verdien 4. Dermed blir alle frekvensene delt inn i hvor høy eller lav frekvensen deres er i forhold til de andre frekvensene til de andre tekstene.

Dersom algoritmen krever det, kan verdiene leses som typer av en klasse og ikke diskrete verdier. Det vil si at de blir ansett som kategorier, hvor 1, 2, 3 og 4 er egne kategori. Ulempen med dette er at variablene mister distansen som viser sammenhengen mellom verdiene, for eksempel at verdien 2 er nærmere verdien 1, enn 4.

Datasett kan også diskretiseres på andre måter enn kvantiler. Frekvensene kan for eksempel deles inn i flere eller færre enn 4 grupper. En annen måte kan være å runde frekvensene opp eller ned, for eksempel at frekvensen 2.321 ble rundet ned til 2.

```
> setwd("/Users/victoriatroland/Desktop/INESS korpus/INESSforsøk/")
> table1 <- read.csv(file="freqtableiness.csv",header=TRUE)
```

```
> table3 <- table1[c(18,19)]
> table2 <- table1[c(2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17)]

> library(dplyr)

> temp1 <- ntile(table2\$CoordTot, 4)
> temp2 <- ntile(table2\$CoordMen, 4)
> temp3 <- ntile(table2\$CoordKomma, 4)
> temp4 <- ntile(table2\$CoordEller, 4)
> temp5 <- ntile(table2\$CoordOg, 4)
> temp6 <- ntile(table2\$Passiv, 4)
> temp7 <- ntile(table2\$Referanse, 4)
> temp8 <- ntile(table2\$AdjType, 4)
> temp9 <- ntile(table2\$Hunkjoenn, 4)
> temp10 <- ntile(table2\$Hankjoenn, 4)
> temp11 <- ntile(table2\$Intetkjoenn, 4)
> temp12 <- ntile(table2\$Genitiv, 4)
> temp13 <- ntile(table2\$Nominativ, 4)
> temp14 <- ntile(table2\$Oblik, 4)
> temp15 <- ntile(table2\$Tall, 4)
> temp16 <- ntile(table2\$Bestemthet, 4)

> table4 <- matrix(c(temp1,temp2,temp3,temp4,temp5,temp6,temp7, temp8,temp9,
temp10, temp11,temp12,temp13,temp14,temp15,temp16),nrow=21,ncol=16)

> data <- data.frame(table4,table1\$ID,table1\$Forfatter)
> write.csv(data, file="quantfreq.csv")
```

Tok bort første kolonne og første rad fra fil, andomiserte deretter linjer og liming av header:

```
cut -d, -f2-19 quantfreq.csv > quantfreq2.csv
tail -n +2 quantfreq2.csv > quantfreq3.csv
gshuf quantfreq3.csv > quantfin.csv
cat header.csv quantfin.csv > wekaquant.csv
```

Eksempel av linje fra ferdig prosessert fil:

```
CoordTot,CoordMen,CoordKomma,CoordEller,CoordOg,Passiv,Referanse,AdjType,
Hunkjoenn,Hankjoenn,Intetkjoenn,Genitiv,Nominativ,Oblik,Tall,Bestemthet,
ID,Forfatter
3,3,1,2,4,2,1,2,2,4,1,1,3,2,3,2,re1,Enger
```

En nærmere forklaring på sammenhengen mellom de forskjellige trekkene, navnene og verdiene kan sees i tabell 5.1 nedenfor. De opprinnelige frekvensene er de samme som jeg kaller kontinuerlige verdier i dette kapittelet.

Trekknr.	Trekk	Kode	Opprinelige frekvenser	Kvantilverdi
1	Koordinering: Totalt	CoordTot	0.048	3
2	Koordinering: men	CoordMen	0.066	3
3	Koordinering: komma	CoordKomma	0.229	1
4	Koordinering: eller	CoordEller	0.032	2
5	Koordinering: og	CoordOg	0.656	4
6	Passiv	Passiv	0.061	2
7	Referanse	Referanse	0.074	1
8	Adjektivtype	AdjType	0.564	2
9	Hunkjønn	Hunkjoenn	0.091	2
10	Hankjønn	Hankjoenn	0.518	4
11	Intetkjønn	Intetkjoenn	0.391	1
12	Genitiv	Genitiv	0.004	1
13	Nominativ	Nominativ	0.485	3
14	Oblik	Oblik	0.511	2
15	Tall	Tall	0.271	3
16	Bestemthet	Bestemthet	2.262	2
17	ID	ID	re1	re1
18	Forfatter	Forfatter	Enger	Enger

Tabell 5.1: Trekk, trekkkoder og verdier med eksempel

5.4 Kvantitativ modellering

I denne delen modelleres datasettet, både det diskrete datasettet og kontinuerlige datasettet, ved hjelp av WEKA. WEKA egnet seg her både på grunn av brukervennligheten og de mange valgmulighetene i preprosessering og i modelleringsfasen. Programmet egnet seg også særlig godt til disse datasettene, da det håndterer både diskrete og numeriske tallverdier automatisk.

WEKA har større valgmuligheter og er bedre egnet enn *Stylo* når det gjelder egne trekk og overvåkede læringsmuligheter.

TiMBL, et lignende maskinlæringsprogram, viste seg å være mindre ideelt for dette forsøket, på grunn av dårlig håndtering av numeriske verdier og få muligheter ved valg av maskinlæringalgoritmer, jfr. 3.3 s. 30.

Tabell 5.4 gir en oversikt over forsøkene og resultatene fra de forskjellige forsøkene.

Etablering av en *baseline*

Ved å bruke den regelbaserte klassifikatoren "ZeroR" etablerte jeg en *baseline* med filen "weka-freq.csv". "ZeroR" velger den mest frekvente klassen i en fil. Forsøket forutså modellen 14.3 % (3 av 21 instanser) riktig, hvor datasettet ble testet på treningssettet. I dette tilfellet var den mest

Tabell 5.2: Tabell med oversikt over forsøk og resultater

Algoritme	Verdier	Resultat (%)	Riktige
Baseline	Kontinuerlig	14.3	3/21
k-NN	Kontiunerlig	62.0	13/21
SVM[1]	Kontinuerlig	66.7	14/21
LMT	Kontinuerlig	38.1	8/21
LADTree	Kontinuerlig	52.4	11/21
k-NN	Diskrete	66.7	14/21
Naive Bayes Multi	Diskrete	76.2	16/21
SVM[2]	Diskrete	66.7	14/21
LMT	Diskret	62.0	13/21
LADTree	Diskret	38.1	8/21

frekvente klassen de 3 instansene av forfatteren "Klippenvåg". Jeg repeterte ikke forsøket med det diskrete datasettet fordi resultatet ville vært det samme.

Forsøk med k-Nearest Neighbour

Ved å bruke en *lazy-learning*-algoritme av typen IB1 på det kontinuerlige datasettet og spesifisere numeriske trekk i Weka klarte modellen å forutsi 62.0 % (13 av 21 instanser) riktig med 10-delt kryssvalidering. Dette er høyere enn *baseline*.

Resultatene viste at forfatterne som var konsekvent riktig klassifisert var: Lindell, Ragde, Oterholm, Kiøsterud, Hamilton. Delvis riktig klassifisert var Hagan (1/2) og Klippenvåg (2/3). Feilklassifisert var Hagan og Lindvåg.

Klassifikatoren IB1 med *lazy-learning* ble brukt på det diskrete datasettet, med 10-delt kryssvalidering. Modellen forutså 14 av 21 riktige instanser, litt bedre enn med kontinuerlige verdier

Riktig klassifiserte forfattere var Johnsgaard, Klippenvåg, Hagan, Oterholm, Kiøsterud og Hamilton. Delvis riktig klassifisert var Ragde (1/2). Feil klassifisert var Enger, Lindvåg og Lindell.

Forsøk med SVM

Det diskrete datasettet ble testet med *Support Vector Machine*. I WEKA ble *Sequential Minimal Optimization* (SMO) valgt, uten normalisering eller standardisering av datane og med en kernel med en eksponent på 2. Det ble brukt en 10-delt kryssvalidering. Modellen forutsa 14/21 riktige. Helt riktig klassifisert var Johnsgaard, Klippenvåg, Hagan, Oterholm, Kiøsterud og Hamilton. Delvis riktig klassifisert var Ragde (1/2). Feilklassifisert var Enger, Lindvåg og Lindell.

Datasettet med kontinuerlige varabler ble også testet på SMO. Etter 4-5 forsøk viste resultatene å optimalisere seg, med en normalisering av dataene og en kernel med eksponent=4.0, på 14/21 riktige. Riktig klassifisert var Lindell, Enger, Ragde, Oterholm og Hamilton. Delvis riktig var Klippenvåg (2/3), Kiøsterud (1/2) og Hagan (1/2). Feilklassifisert var Johnsgaard og Lindvåg.

Forsøk med Naive Bayes

I dette forsøket ble det diskrete datasettet testet på en *Naive Bayes Multinomial*-modell (McCallum et al., 1998). Modellen forutså 76.2 % eller 16/21 riktige instanser. Riktig klassifisert var Johnsgaard, Klippenvåg, Hagan, Oterholm, Kiøsterud og Hamilton. Delvis riktig var Lindvåg (1/2), Ragde (1/2) og Lindell (1/2). Kun Enger ble feilklassifisert.

NBM viste seg å ikke være egnet til det kontinuerlige datasettet. Da det kontinuerlige datasettet ble testet med NBM, men presterte settet bare 1/21 riktige utfall og ble dermed tatt ut.

Forsøk med *decision trees*

I WEKA (versjon 3.6.12) finns det 16 mulige algoritmer som bygger trær. For det diskrete datasettet er *functional trees*, J48 og J48graft (danner C4.5-trær), LADTree, *logistic model trees* (LMT), REPtree, NaiveBayes-tree, *RandomForest* og *RandomTree*. På grunn av de numeriske verdiene og at det finnes flere enn to klasser passet FT, LADTree og LMT best til datasettet.

Resultatene av det diskret datasettet og træralgoritmene var: FT: 0/21, LADTree: 8/21, LMT: 8/21

Resultatene av det ikke-diskret datasettet og træralgoritmene var: FT: 0/21, LADTree: 11/21,LMT: 13/21

FT klassifiserte konsekvent alle instansene i den mest frekvente klassen “Klippenvåg”, lik “ZeroR”, og ble derfor tatt ut av forsøket. Modellen tilførte ikke noe nytt til forsøkene og datasettene lot til å være uegnet til FT.

LADTree med diskrete verdier ble Lindvåg og Klippenvåg riktig klassifisert. Delvis riktig var Oterholm (1/2), Kiøsterud (1/2) og Hamilton (1/2). Feilklassifisert var Enger, Johnsgaard, Hagan, Ragde og Lindell.

Av de kontinuerlige forsøkene hadde LADTree 11/21 og Oterholm og Hamilton var riktig klassifisert. Delvis riktig klassifiserte var Lindell (1/2), Ragde (1/2), Hagan (1/2), Kiøsterud (1/2), Lindvåg (1/2) og Klippenvåg (2/3).

LMT med diskrete verdier hadde 13/21 riktige utfall. Riktig klassifisert var Johnsgaard, Klippenvåg, Hagan, Oterholm, Kiøsterud og Hamilton. Ingen ble delvis klassifisert. Feilklassifisert ble Enger, Lindvåg, Ragde og Lindell.

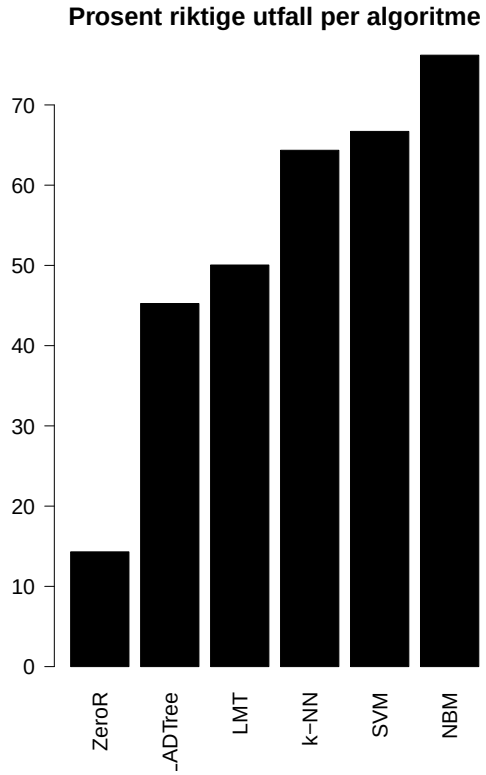
LMT med kontinuerlige verdier hadde 8/21 riktige utfall. Riktig klassifisert ble Hamilton og Klippenvåg. Delvis riktig klassifisert ble Lindell (1/2), Oterholm (1/2) og Kiøsterud (1/2). Feilklassifiserte var Johnsgaad, Enger, Ragde, Hagan, og Lindvåg.

5.5 Resultat: funn, diskusjon og konklusjon

Resultatene i tabell 5.4 viser at alle modellene utfører forsøkene bedre enn *baseline*. Enkelte modeller forutsa forfatterskap bedre enn andre.

Resultatene viser at alle algoritmene klarer å forutse riktig utfall i mye større grad enn *baseline*. Den modellen som klarer å forutse flest riktige utfall av forfatterskap er Naive Bayes Multinomial

med de diskrete verdiene. Ellers presterer både SVM og k-NN godt for begge datasettene. LMT og LAD varierte mellom datasettene. Algoritmene fungerer jevnt over best på diskrete verdier enn kontinuerlige verdier.



Figur 5.3: Algoritmer og riktige utfall sammenlagt

Figur 5.3 illustrerer hvilke algoritmer som prosentvis i økende grad forutsier riktige utfall. Lavest er baselinen, LMT og LADTree. Høyest er k-NN, SVM og NBM. Krysningstabell 5.3 med antall riktige og uriktige antagelser gjort av alle modellene til sammen gir en oversikt over klassifiseringene som ble gjort[3]. Det er mulig å se tendenser i algoritmene. Krysningstabellen viser at forfatterne Klippenvåg (24/27)[4], Oterholm (16/18), Hamilton (17/18), Hagan (11/18) og Kiøsterud (14/18). Forfatterne Enger (2/18), Lindvåg (4/18), Lindell (7/18), Ragde (8/18) og Johnsgaard (8/18) hadde høyest feilklassifisering tilsammen av alle algoritmene.

Tendensene i tabellen avslører at enkelte forfattere er enklere for algoritmene å forutsi enn andre. Man ser også at andre forfattere systematisk trekker presisjonen ned.

Dette kan skyldes menneskelig feil, men frekvensene og forfatterne ble undersøkt opp mot hverandre i filene i ettertid. Trolig ligger frekvensene til en av tekstene til Lindell tett nok opp til en av Enger sine tekster at de blir forvekslet.

Klippenvåg ble oftest klassifisert riktig. Trolig fordi Klippenvåg har flere tekster i korpuset enn de andre forfatterne og derfor bli lettere å forutse.

[3] Baseline-forsøket er ikke tatt med

[4] Tallene etter forfatterne står for riktig klassifisering i forhold til antall klassifiseringer.

Tabell 5.3: Figur med antall riktige og uriktige forfatter i CM av alle forsøkene

re	el	mj	ok	ph	ar	ao	ul	ek	dh	
2	1	2	1		1		10		1	re = Enger
	4			1	11	2				el = Lindvåg
7	1	8	1			1				mj = Johnsgaard
3			24							ok = Klippenvåg
	1	1		11			5			ph = Hagan
1	6				8	1		2		ar = Ragde
		1				16			1	ao = Oterholm
10	1						7			ul = Lindell
					3		1	14		ek = Kiøsterud
		1							17	dh = Hamilton

5.5.1 Effektiviteten til trekkene

Trekkene ble vurdert med *Information Gain* (IG) i WEKA, med *Ranker* og 10 delt kryssvalidering. Både de kontinuerlige verdiene og diskrete verdiene ble vurdert. I figur 5.5 er tallene fra figur 5.4 lagt sammen på s. 66.

IG tar hvert trekk hver for seg og estimerer hvor mye informasjon trekket gir oss om riktig klassifisering (Daelemans and Van den Bosch, 2005, s. 29f).

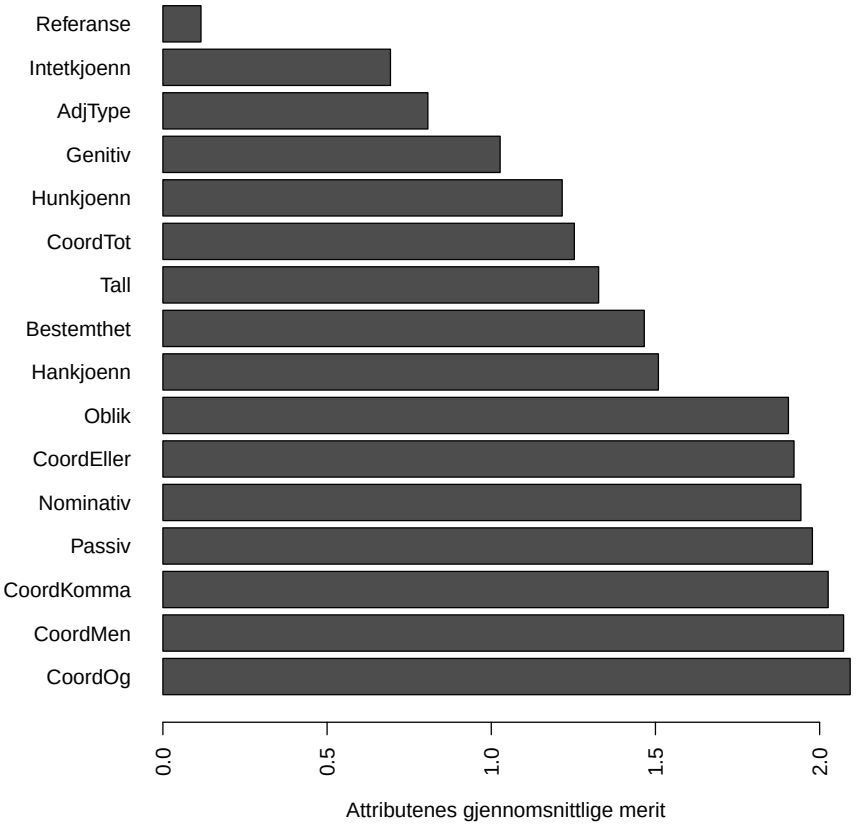
Hver for seg viser figuren 5.4 at begge typene trekk overlapper til en viss grad. I tabell 5.4 viser det seg at koordineringskontruksjonene (unntatt CoordTot), passiv, nominativ og oblik presterte høyest når det kom til nyttig informasjon trekket ga modellene til sammen.

Mellom de diskrete verdiene og de kontinuerlige verdiene var IG-verdiene ikke svært ulike. De fleste trekkene ligger tilnærmet i samme område som i figur 5.4. Den største forskjellene i posisjon er trekkene CoordOg, CoordMen og Hunkjoenn. Verdimessig varierer de to første mellom datasettene. I tillegg har alle trekkene for de kontinuerlige verdiene jevnt over høyere verdi enn trekkene for de diskrete verdiene. Til tross for dette presterte datasettet med de diskrete verdier litt bedre i algoritmene.

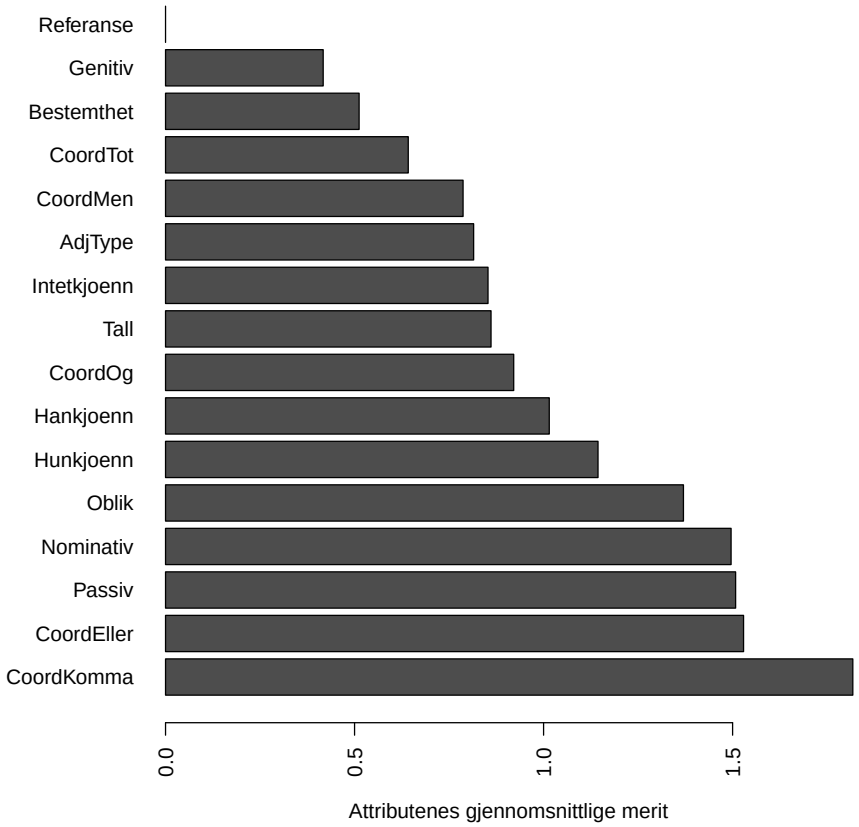
Påfallende er at figuren i tabell 5.4 for diskrete verdier har et trekk som har 0 i IG. Trekkene for det diskrete datasettet lavere IG-verdier enn i det kontinuerlige datasettet. Til tross for dette ser det ikke ut til å påvirke resultatene med de forskjellige algoritmene dersom man sammenligner resultatene i tabell 5.4.

De logikkbaserte algoritmene, LMT og LADTree var forventet å prestere best på det diskrete datasettet. Det presterte jevnt over best, men det kontinuerlige settet presterte best på LADTree. Det er uvisst hvorfor, men kan tenkes at LADTree presterer bedre med variabler som inneholder mer informasjon enn det diskrete datasettet og er sensitiv til kontinuerlige verdier.

Tabellen 5.4 gir informasjon om resultatene med standardavvik har gitt om trekkenes IG for begge datasettene. *Meritt* sier noe om hvor mye informasjon et trekk gir. *Rang* indikerer intervallet for rangeringen i de forskjellige forsøkene.

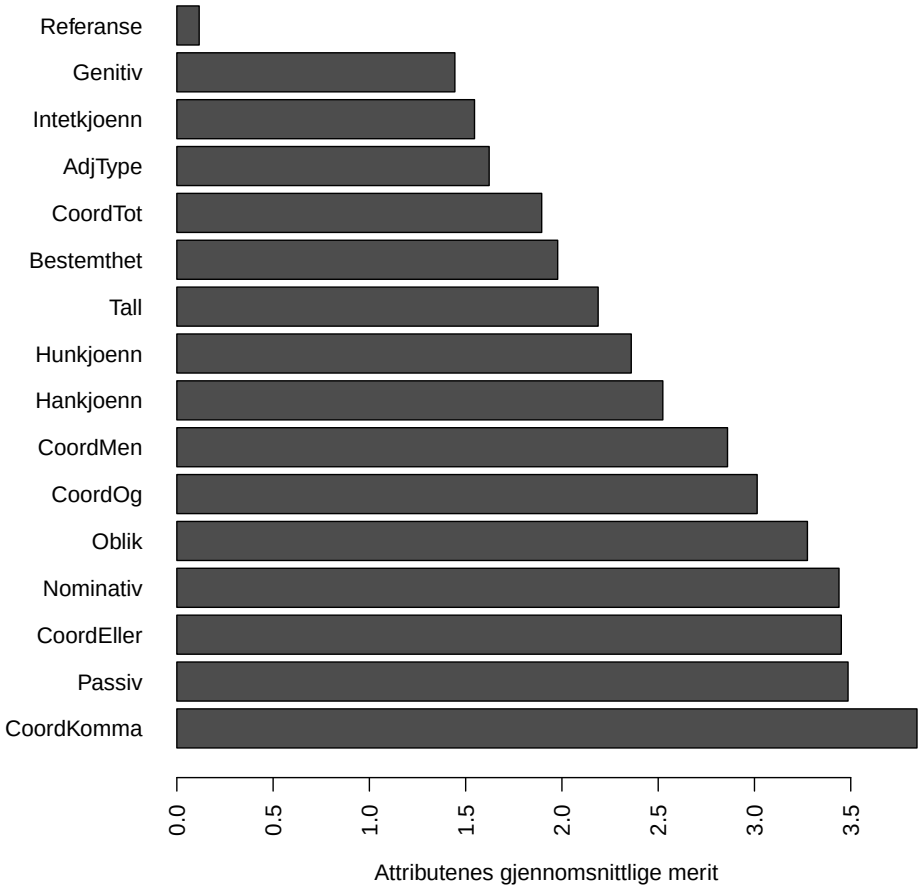


(a) Information Gain av trekk for kontinuerlige verdier



(b) Information Gain av trekk for diskrete verdier

Figur 5.4: Informatio Gain



Figur 5.5: Sammenlagt Information Gain med diskrete og kontinuerlige verdier

	meritt	+- meritt	rang	+- rang
CoordKomma	3.84	0.36	5.70	3.48
Passiv	3.49	0.53	9.70	5.31
CoordEller	3.45	0.45	9.80	4.08
Nominativ	3.44	0.49	10.00	4.40
Oblik	3.28	0.70	11.30	7.11
CoordOg	3.01	0.61	13.10	3.73
CoordMen	2.86	0.92	12.40	6.38
Hankjoenn	2.52	0.50	17.70	4.05
Hunkjoenn	2.36	0.87	19.00	6.28
Tall	2.19	0.73	20.90	4.73
Bestemthet	1.98	0.87	21.40	4.95
CoordTot	1.90	1.50	21.10	9.72
AdjType	1.62	1.24	22.80	6.27
Intetkjoenn	1.55	1.17	22.80	7.87
Genitiv	1.44	0.66	25.20	3.62
Referanse	0.12	0.35	29.10	2.11

Tabell 5.4: Attributtenes meritter og rangering

5.5.2 Dokumentasjon av forsøkene og resultatene

På grunn av størrelsen av dokumentasjonen av forsøkene er deler lastet opp på internett og ikke vedlagt i teksten. Resultatene i vedlegg K (s. 125) viser kun oppsummeringen av forsøkene. INESS-dokumentasjonen som er lastet opp på internett inneholder de fullstendige WEKA-genererte resultatene. Dokumentasjonen er tilgjengelig via linken:

<figshare.com/s/446652b6f8c211e4bc2106ec4b8d1f61>.

5.5.3 Diskusjon og konklusjon

Diskusjon

I kapittel 2 nevnes det NB generelt presterer bedre på diskrete trekk. I dette forsøket ble NB testet på det kontinuerlige datasettet, men ble ekskludert i ettertid for å være en uegnet algoritme til datasettet. NBM fungerte kun på det diskrete datasettet, som tolket verdiene som kategorier og hadde lite treningsdata. NBM presterte best av alle modellene med 76.2 % riktige utfall. Dette viser at det NBM var godt egnet til forsøket og at det diskrete datasettet kan fungere tilfredsstillende med datasett som er tolket som kategorier.

SVM presterte like høyt på det diskrete datasettet og det kontinuerlige datasettet, men presterer generelt bedre med kontinuerlige verdier (Kotsiantis, 2007). Siden SVM krever en større mengde treningsdata for å oppnå optimal bestemmelsesrate, enn NB (Kotsiantis, 2007), ville trolig SVM-forsøkene prestert høyere dersom datasettene hadde vært større. Videre ville trolig det kontinuerlige datasettet prestert bedre enn det kontinuerlige.

Valget av diskretisering kunne vært gjort annerledes. Målet med diskretisering var å komprimere informasjonen og omgjøre verdiene i datasettet til verdier som kunne behandles kategorisk eller diskret. Forskning på sin side har indikert at en diskretisering basert på gevinstene ved entropi og distribusjonen av potensielle inndelinger i et sett kan øke prestasjonen ved maskinlæring (An and Cercone, 1999).

I tabell 5.3 ble både Lindell og Enger forvekslet med hverandre i 10/18 ganger. Dette kunne skyldes at frekvensene til en eller begge av hver forfatters tekst ligger nært hverandre og er derfor blitt forvekslet eller en menneskelig feil. Det samme gjelder Klippenvåg. Når Klippenvåg blir feilklassifisert, blir han konsekvent klassifisert som Enger.

Siden INESS-tekstene ikke kunne kontrolleres for lengde, kunne det påvirket frekvensene. Eksempelvis, i Kiøsteruds "Ved fjellets fot" forekom ingen passive setninger i INESS-søket. Dette gjør at frekvensen ikke er målbar for denne teksten og kan ha påvirket resultatene i maskinlæringsdelen. Til tross for null forekomster av passiv i "Ved fjellets fot" valgte jeg å inkludere trekket fordi det kunne vært karakteristisk for andre tekster og derfor nyttig å inkludere. I 5.5 bekreftes det at passiv kan brukes til å indikere forfatter. Passiv er trekket med nest høyest IG av alle de andre.

På forhånd antok jeg at hunkjønn-trekket kom til å være mer karakteristisk enn hva IG-tabellen 5.5 viser. I tabell 5.4 er hunkjønn med diskrete verdier høyere enn med de kontinuerlige verdiene. Det indikerer at forfatterens karakteristiske frekvens uttrykkes bedre med diskrete verdier i tilfellet

med hunkjønn. Dette kan være fordi de kontinuerlige frekvensene ligger nært hverandre på tvers av forfattere og en diskretisering av verdiene tydeligere skiller frekvensene fra hverandre.

Koordineringstrekkene viste seg å være effektive trekk. Det var variasjon av IG-verdier innad settene, men i figur 5.5 ligger alle koordineringsformene, untatt KoordTot, i den øvre halvdelen av skalaen. Et nærmere blikk på frekvensene i tabell I.1 indikerer at valg av koordingeringsform er subjektivt. Der vises en tendens til at den mest frekvente koordineringsformen er "og", for de fleste forfatterne. Unntakene er Klippenvåg og Kiøsterud. Lindell og Lindvåg er forfatterne med høyest verdier av "eller" i sine tekster. Slike åpenbare tendenser kan delvis forklare hvorfor koordineringstrekkene har relativt høy IG-verdier. Derimot forklarer det ikke hvorfor KoordTot har en relativt lav IG-verdi i forhold til de andre trekkene. En potensiell forklaring er at KoordTot ikke er relativisert med en absolutt verdi: Totalt antall ord, og dermed ikke fått et representativt uttrykk til tekstene. En annen forklaring er at totalt antall koordineringsformer ikke er karakteristisk for forfattere, men at hvilke koordineringsformer er viktigere for å skille forfattere fra hverandre.

Referanse viste seg å være et trekk med lav IG-verdi. Det er uklart hvorfor referanse ikke er et effektivt trekk. En forklaring er at forfattere eller prosatekster generelt har lik grad av referering i diskursen. En annen forklaring er at referanse ikke er gjennomgående annotert for i INESS og skaper uklare frekvenser.

På forhånd var tall antatt å være et innholdsavhengig trekk og dermed mindre effektivt for å skille forfattere fra hverandre enn eksempelvis innholdsfrie trekk som koordinering. IG-verdiene indikerer at trekket presterer middelmådig i forhold til de andre trekkene i datasettet. En forklaring på dette kan være at forfattere har en underbevisst tall-preferanse. På en annen side kan IG-verdiene indikere at forfatternes bøker inneholder tematiske likheter, siden tema ikke ble kontrollert for.

Adjektivtype var på forhånd antatt å ha høyere påvirkning av klassifiseringsraten enn den tilsynelatende har. Trekket kan tyde på er i mindre en forfatterpreferanse enn tidligere antatt i studien. Det er mulig adjektivstilling ikke er like utelukkende valgfritt som tidligere antatt, men er påvirket av andre motiver, som fokus og beskrivelsesform i setninger.

Kasus var på forhånd antatt å oblik ville være trekket med høyest påvirkning klassifiseringsraten, siden oblik ikke er en obligatorisk del av en setning, men avhengig av predikatet. Oblik fikk en relativt høy IG-verdi og er et av trekkene som presterer høyest av trekkene ekstrahert fra INESS. Genitiv var på forhånd ikke antatt å ha være svært informativt, siden det er innholdsavhengig. Nominativ var heller ikke forventet å ha høy IG, fordi det var antatt at siden nominativ var obligatorisk i en setning, ville det være lite indikativt av forfatterskap. Siden passiv har vist seg å være svært indikativt på forfatterskap og passiv fjernes agens, som ofte er subjektet kan de ha en sammenheng. Subjektet korrelerer med nominativ, som kan tenkes at er sammenhengen mellom nominativ og passiv. Noe som kan forklare den høye IG-verdien til nominativ, siden passiv har høy verdi. På en annen side kan det tenkes at siden nominativ er høyfrekvent og er relativisert ved hjelp av de andre kasusene at frekvensen uttrykker enten en tilstedeværelse eller et fravær av de andre kasusene.

Bestemthet var på forhånd antatt å være diskursavhengig, som gjør det potensielt innholdsavhengig. Bestemthet viste seg å ha en lav til middelmådig påvirkning på klassifiseringsraten, i forhold til de andre trekkene. Den lave prestasjonen kan indikere at bestemthet er mer innholdsavhengig er

tidligere antatt. På en annen side kan de tenkes at bestemthet ikke er svært indikativt på forfatterskap.

Trekkene hentet fra INESS er gjennom studien beskrevet som syntaktiske trekk. Dette er en uriktig og upresis benevning. Referanse er diskursavhengig, som derfor kan klassifiseres som et pragmatisk trekk. Kasus uttrykker semantiske roller og kan dermed klassifiseres som semantiske trekk. Passiv kan være semantisk motivert ved å fjerning av agens, men påvirker syntaksen til en setning. Tall, bestemthet, kjønn og adjektivstilling kan sies å være morfosyntaktiske siden de kan være eller påvirke formen av en hovednode og de andre nodene i frasen. Koordineringer kan sies å være syntaktiske, siden de styrer forholdene mellom to ledd eller setninger. Benevningen "syntaktisk" ble brukt fordi INESS er en syntaktisk trebank og målet var å ekstrahere ut trekk fra trebanken, som på forhånd var antatt å være primært syntaktiske. Det at trekk de valgte trekkene i ettertid ikke var rent syntaktiske er ikke viktig siden målet var ekstraheringen av trekk i INESS og ikke syntaktiske trekk i seg selv. Det er likevel viktig å presisere at benevningen som har vært brukt tidligere om trekkene kan ha vært upresis.

Generelt hadde det kontinuerlige datasettet høyere IG-verdier. Til tross for dette presterte algoritmer med det diskrete datasettet bedre enn det kontinuerlige datasettet. En forklaring er at enkelte av modellene er mindre egnet til kontinuerlige verdier enn diskrete datasett, som NBM. En annen forklaring kan være at IG-verdiene til de diskrete settene reflekterte nytten til trekkene mer optimalt enn det kontinuerlige settet. Til sist kan datasettene ikke ha vært store nok til å kunne oppdage ordentlige tendenser i prestasjonene til datasettene. En annen mulighet er at variansen innen IG (meritt) er større for det kontinuerlige settet og dermed påvirker presisjonen til modellene negativt.

Tekstfrekvensene ble hentet ut manuelt og plassert i en .csv-fil. Dette var problematisk, fordi det øker muligheten for menneskelig feil. En automatisk innhenting av frekvenser hadde vært optimalt, men er ennå ikke en mulighet i INESS-trebanken. I senere forsøk burde dette implementeres.

Det var usikkerhet om betydningen av den andre kolonnen til *merit* i forsøkene med IG. Grunnen til dette var manglende dokumentasjon i WEKA. Trolig indikerer kolonnen standardavviket til resultatet av kryssvalideringen. Et forsøk uten kryssvalidering gir ikke en *merit*-kolonne i WEKA.

Jeg opplevde tidvis problemer ved søk i INESS-trebanken. Blant annet ble indekseringen av tekstene tidvis uriktig. Når systemfeil, som indekseringen ble oppdaget ble de rettet opp igjen i ettertid av systemansvarlig. Det er likevel mulighet for at det har forekommet feil av den grunn, som ikke er registrert i studien.

Konklusjon

Trekkene hentet ut fra INESS kunne anvendes til forfatterattribuering ved hjelp av maskinlæring, med en høyeste klassifiseringsrate av forfattere på 76.2 %. Et datasett som diskretiserte frekvensene ga høyere riktig klassifisering enn det kontinuerlige settet. NBM maskinlæringalgoritmen som hadde høyest grad av riktig klassifisering. Trekkene hadde forskjellige IG og de fire med høyest IG var de koordinerende trekkene, passiv, nominativ og oblik. Referanse hadde lavest IG.

Kapittel 6

Diskusjon og konklusjon

I denne delen blir funnene fra kapittel 4 og 5 drøftet sammen. Først presenteres hovedfunnene, deretter diskuteres funnene og utformingen av studien. Til sist presenteres en konklusjon og forslag til videre forskning.

6.1 Oversikt over hovedfunnene

I kapittel 4 ble leksikalske trekk anvendt og hadde tilfredsstillende resultater i forsøk med forfatterattribuering. Statistiske metoder, som MDS, *Cluster analysis* og *Consensus tree*, ble anvendt og kunne i noe grad skille forfattere fra hverandre, avhengig av hvilke parametre som ble anvendt. Forsøk med MFW-parameteret presterte høyere enn forsøk med andre parametre som for eksempel *culling* og *sampling*. Noen forfattere ble oftere riktig gruppert enn andre, eksempelvis Oterholm. Andre forfattere ble oftere gruppert feil, som Enger. Engers feilgruppering kan skyldes lite samsvar mellom de leksikalske trekkenes frekvenser på tvers av eller innad tekstene, som videre indikerer at trekkene ikke er like effektive på enkelte forfattere som de er på andre.

I kapittel 5 ble en rekke trekk ekstrahert fra INESS. Koordinering, passiv, kasus og noen utvalgte grammatiske kategorier viste seg å være effektive trekk i forfatterattribueringsforsøk med maskinlæring. Diskretisering av datasettet sammen med NBM presterte høyest, med 76.2 % riktig klassifisering av forfatterne ved hjelp av INESS-trekkene.

Utvalget av korpustekstene viste seg å være påvirket av lesbarhetsgrad, som igjen påvirket parsingsgrad i INESS. Liks er basert på gjennomsnittlig setningslenge og lange ord, som trolig utgjør grunnlaget for sammenhengen mellom Liks og parsingsgrad. Med andre ord kan lange setninger påvirke Liks-skalaen oppover og dermed gjøre setninger vanskeligere å parse i INESS.

6.2 Vurdering av utforming og utførelse av studien

Hovedforskjellene mellom *Stylo*- og INESS-forsøkene er ulik grad av automatikk i prosessen og trekktyper. I INESS-forsøkene måtte deler av metoden utvikles og trekkene preprosesseres for å kunne utføre forsøkene i WEKA. I INESS ble nye trekk ekstrahert fra trebanken, i motsetning til i

Stylo-forsøkene hvor programmet automatisk ekstraherte trekkene.

Siden det ikke finnes en mal for utførelse av forfatterattribueringsforsøk, er teksten preget av en metodisk, men eksperimentell tilnærming til de lingvistiske trekkene, korpusdannelsen og de statistiske metodene. Valgene som er tatt er begrunnet og dokumentert utførlig for å tillate reproduksjon.

De statistiske metodene som er valgt er metoder som er anvendt i tidligere stilometriske undersøkelser. Bruk av 10-delt kryssvalidering var en forsikring om stabilitet av resultatene i modellen i kapittel 5. Resultatene av de statistiske analysemetodene i *Stylo*-forsøkene var åpen for subjektiv tolkning, mens forsøkene i INESS resulterte i et numerisk tall for riktig klassifisering.

Trekkvalgene og programmene påvirket utformingen av eksperimentene ved at studien ble delt inn i to forskjellige understudier hvor ulike trekk og statistiske metoder ble benyttet. Forskjellige statistiske metoder gjør det vanskeligere å sammenligne resultatene på tvers av *Stylo* og INESS og vurdere dem opp mot hverandre. Eksempelvis har både *Stylo* og WEKA maskinlæringsalgoritmer, men de ulike utformingene av disse gjør det utfordrende å sammenligne resultatene av de to forsøkene. En optimal løsning ville ha vært å gjennomføre forsøk på datasettene i samme program, noe som ikke ble gjort her.

Frekvenser av trekk er den kvantitative målestokken i stilometriske undersøkelser. I undersøkelsene med *Stylo* ble n-gram-frekvenser automatisk relativisert i forhold alle n-grammene. Frekvensene i INESS var mer krevende å relativisere, fordi prosessen ikke var automatisk og frekvensene måtte relativiseres i forhold til andre sammenlignbare og relative frekvenser for å oppnå representativitet.

For å minimere menneskelige feil er prosessen for å ekstrahere trekk, prosessere tekster og anvende de statistiske metodene er gjort automatisk i den grad det er mulig. Ekstrahering av INESS-frekvensene og kategoriseringen av tekstene til *Stylo* var de delene av prosessen som måtte utføres manuelt. En fremtidig studie bør automatisere ekstraheringen av INESS-frekvensene.

Både *Stylo*-forsøkene og INESS-forsøkene ga tidvis sammenfallende resultater av forfattere i undersøkelsene. Forfatterne som ble gruppert feil eller feilklassifisert var ofte de samme i begge forsøkene. Forfatteren Enger viste seg på tvers av undersøkelsene å være vanskelig å klassifisere ved hjelp av trekkene og de statistiske modellene som ble valgt. En nærmere undersøkelse av tekstene ville trolig kunne gi en indikasjon på hvorfor dette er tilfellet.

6.2.1 Samsvar med annen forskning

Tidligere forskning har indikert at forsøk med ordbaserte og syntaktiske trekk sammen har vært effektive i forfatteranalyser. I denne studien ble ikke leksikalske og syntaktiske trekk vurdert sammen, men gitt tidligere forskning ville et forsøk med begge typer trekk potensielt kunne øke presisjonene til modellene.

Prestasjonene med maskinlæringsalgoritmene oppfylte delvis tidligere forventninger basert på forskning. NBM hadde som forventet høyest klassifiseringsrate med et relativ lite treningssett og diskrete trekk. SVM hadde dårligere resultat enn NBM. Dette var ikke overraskende, siden SVM presterer bedre gitt høyere dimensjonalitet i treningssettet. k-NN var forventet å presterte bedre med

det kontinuerlige datasettetet enn det diskrete settet, ikke var tilfelle i dette forsøket. Trolig ble det diskrete datasettet tolket som kategorier eller naturlige tall, noe som kan ha økt klassifiseringsratioen. LADTree og LMT presterte adekvat, men dårligere enn de andre algoritmene, selv om det var forventet at de presterte lavere enn enn SVM. Deler av den relativt lave presisjonene kan komme av irrelevante attributter.

Den statistiske analysemetoden PCA er tidligere blitt kritisert for ikke å prestere tilfredsstillende i forsøk med ikke-profesjonelle forfattere sammenlignet med forsøk med profesjonelle forfattere. Denne studiens korpus inneholdt kun tekster skrevet av profesjonelle forfattere, og for å unngå en potensiell overvurdering av forskjellene mellom skrivestilene ble PCA ikke brukt i denne studien. Studien anvendte andre statistiske analysemetoder. Analysene grupperte forfattere med en adekvat grad av sammenfall, men en nærmere studie av effekten av forskjellige statistiske metoder i forbindelse med forfatterattribuering ville ha vært ønskelig.

Det finnes foreløpig ikke konsensus rundt et felles rammeverk for forfatterattribuering (Rudman, 1998). Med fordel kunne dette gjøres ved både trekkvalg og statistiske analyser i større grad enn tidligere. Dette kan av flere grunner være en komplisert oppgave. Eksempelvis er enkelte trekk, som skrivekontroll og parsere, språkspesifikke eller krever språkspesifikke verktøy. Problemstillingene som er valgt vil også innvirke på dannelsen og annotasjonen av korpus. Statistiske metoder velges etter funksjon, for eksempel hvorvidt man ønsker å klassifisere etter et gitt skjema eller oppdage en underliggende struktur.

6.2.2 Korpuset og trekkvalg - hvor egnet var disse?

Korpuset ble dannet med mål om å undersøke forfatterattribuering. Utvalget var begrenset til tekster i INESS som oppfylte parsingsgrensen. Dette påvirket antall forfattere og tekster som kunne inkluderes, lesbarhetsgraden og sjanger (prosa). Dette førte til at det ikke kunne kontrolleres for enkelte faktorer, som tema og alder. Korpuset kontrollert for utgivelsesår, antall tekster og kjønn.

I kapittel 3 ble forholdet mellom lesbarhet og parsingsgrad undersøkt. De viste en korrelasjon mellom faktorene som impliserte en påvirkning av utvalget av tekster fra INESS. Dette kan ha ført til at tekster av samme forfatter har blitt utelatt, og gjort samlingen av tekstene til forfatterne mer homogen enn hva som egentlig er representativt for forfatterne. Homogeniteten kan igjen ha påvirket frekvensene av trekkene og utfallene av de statistiske analysene. Lesbarheten kan potensielt sees som en egenskap av forfatteren. Trolig finnes det en korrelasjon mellom faktorene, f. eks. har Ragde sine tekster en Liks-verdi på 23 og 24 og Oterholm en Liks-verdi på 19 og 18. Dersom det finnes en korrelasjon ville dette kunne påvirke undersøkelsene, blant annet INESS-frekvensene, da disse var valgt ut fra parsingsgrad. Frekvensene som ble hentet ut kunne ha vært mer representative for tekster med høy grad av parsing.

Ved nedlasting av tekstene ble ikke det originale tekstoppsettet beholdt. Det utelukket enkelte applikasjonsspesifikke trekk i studien. En anvendelse av applikasjonsspesifikke trekk ville ha vært interessant som supplement til de andre trekkene i studien.

Siden korpuset besto av 10 forfattere og 2-3 tekster per forfatter var det mest egnet til forfat-

terattribuering. Forsøket med "Kjønnskorpuset" viste at det var lite egnet til forfatterprofilering av egenskaper som kjønn.

Forsøk mellom "Novellekorpus" og "Likelangtkorpus" indikerer at tekstlengde kan ha betydning for resultatene. Jo lengre tekstene er, jo enklere er det for de statistiske modellene å gruppere forfattere sammen. Dette kan skyldes høyere eller mer karakteristisk representativitet av trekkene med økt tekstlengde.

6.2.3 Programmene - hvor egnet var de?

Stylo var et nyttig verktøy til forsøk med leksikalske, og tegnbaserte trekk og var programmert for denne typen trekk. *Stylo* hadde en grafisk *interface* og var enkel å bruke. Mulighetene for manipulering av ordlister og statistiske modeller gjorde *Stylo* godt egnet til å utføre forfatterattribueringsforsøk. Det var begrensninger ved valg av maskinlæirngsalgorimer, noe som gjorde programmet mindre optimalt enn WEKA til forsøk med maskinlæring. *Stylo* er programmert primært for undersøkelser med leksikalske trekk, som programmet automatisk ekstraherer. En mer fleksibel trekkbehandling og maskinlæringsmuligheter ville ha gjort programmet mer brukervennlig.

WEKA er primært laget for maskinlæring og ikke for stilometriske undersøkelser. Preprosessering av tekstene gjorde WEKA forberedelsesmessig mer krevende å bruke enn *Stylo*. Muligheten til å fjerne trekk, som ID, før klassifiseringsdelen var nyttig. WEKA hadde et grafisk *interface* som gjorde programmet enkelt å bruke. Den eneste ulempen er at den grafiske *interfacen* kun er egnet for mindre datasett. Valgmulighetene for maskinlæringsalgoritmene og kryssvalideringen gjorde WEKA optimalt til INESS-undersøkelsene, i forhold til andre program som TiMBL og *Stylo*. Resultatene ble generert i en krysstabell og tabell med presisjonsfeilene i forsøket. Muligheten for å undersøke IG var hensiktsmessig for undersøke nærmere hvilke trekk modellene opplevde som nyttige. Mulighet for å måle effektiviteten av trekkene var spesielt viktig for trekk som tidligere ikke har vært teset på norske tekste. Begrensningen til WEKA var en lengre og mer omfattende preprosesseringsfase for å klarkjøre datasettene.

6.3 Konklusjon

> *"Some of these days spurious writings will be detected by this test. Mind, I told you so."*
>
> — Augustus De Morgan, *Memoir of Augustus De Morgan*

Hypotesen i studien var at stilometriske metoder kunne anvendes til å gjenkjenne forfattere via norske prosatekster. Funnene viser at en rekke stilometriske metoder kan gi en statistisk sannsynlig attribuering av forfattere av norske prosatekster.

Både leksikalske og syntaktiske trekk ble ekstrahert. Trekkene kunne anvendes til å klassifisere og gruppere forfattere riktig, påvirket av modellens innstillinger og datasett. Både statistiske analyser og maskinlæring kunne gjennomgående gruppere og klassifisere forfattere i forsøkene som ble utført. NBM, SVM og k-NN presterte best av maskinlæringsalgoritmene. Av de statistiske analysemetodene presterte alle adekvat, spesielt med manipulering av ordlisten (MFW), men på grunn av inherente egenskaper måtte resultatene tolkes mer subjektivt enn i INESS-forsøkene. I både *Stylo*-forsøkene og INESS-forsøkene var den oftest feilgrupperte forfatteren den samme.

INESS viste seg å være et fleksibelt verktøy for ekstrahering av trekk. Trekkene var annotert for dyp syntaktisk analyse og tekstene kunne lastes ned for å ekstrahere andre trekk.

Samlet kunne informasjonen i trekkene fra INESS-trebanken være delaktig i å forutsi forfatterskap, høyere enn *baseline* i alle forsøkene, med beste resultat på 76.2 %. Enkelte trekk hadde høyere verdi for modellene enn andre trekk, som koordinering, passiv og oblik. Referanse var det trekket som presterte dårligst.

N-gram-trekkene, både ordgram og bokstavgram, kunne gruppere ut fra forfattere, men ikke kjønn. På grunn av et lite egnet korpus, kunne de ikke klassifisere ut fra kjønn.

Lesbarhet og parsingsgrad er faktorer som ville kunne være interessant å undersøke nærmere. De har påvirket undersøkelsene i utvalg av forfatterne og muligens frekvensene hentet fra INESS.

6.4 Videre forskning

De lovende resultatene for forfatterattribuering av norske tekster i denne studien åpner for mer stilometrisk forskning på norske tekster, deriblant mer språkspesifikk forskning. I denne studien ble grammatisk kjønn undersøkt motivert ut fra en potensiell subnormering. Videre forskning åpner for undersøkelser av andre trekk påvirket av norsk variasjon og subnormering, eksempelvis a- og en-endelser i bokmål.

Stylo viste seg å være et nyttig verktøy i gruppering av forfattere. Dypere undersøkelser av leksikalske trekk og frekvenser kan gi en indikasjon på optimalisering av trekkparametre, som n og MFW, i lignende forfatterattribueringsundersøkelser. Det er mulig de optimale parametrene for norske tekster avviker fra på andre språk.

De syntaktiske trekkene som er hentet ut er en liten andel av de trekk som potensielt kan hentes ut i INESS. Videre forskning kan anvende INESS til å ekstrahere andre trekk, eksempelvis omskrivningsregler. Den samme fremgangsmåten som er utviklet her kan anvendes.

Videre stilometriske undersøkelser bør etter hvert føre til utvikling av et allment akseptert felles rammeverk. Dette innbefatter enighet om begrep og metoder. Til nå har det vært lite konsensus om metodologi innen fagdisiplinen.

Bibliografi

Abbasi, A. and H. Chen (2008). Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. *ACM Transactions on Information Systems (TOIS) 26*(2), 7:1–7:29. 7

An, A. and N. Cercone (1999). Discretization of Continuous Attributes for Learning Classification Rules. In N. Zhong and L. Zhou (Eds.), *Methodologies for Knowledge Discovery and Data Mining*, pp. 509–514. Springer Berlin Heidelberg. 67

Argamon, S., M. Koppel, J. Fine, and A. R. Shimoni (2003, January). Gender, Genre, and Writing Style in Formal Written Texts. *Text - Interdisciplinary Journal for the Study of Discourse 23*(3), 321–346. 6

Argamon, S., C. Whitelaw, P. Chase, S. R. Hota, N. Garg, and S. Levitan (2007, April). Stylistic Text Classification Using Functional Lexical Features. *Journal of the American Society for Information Science and Technology 58*(6), 802–822. 9

Baayen, H., H. van Halteren, A. Neijt, and F. Tweedie (2002). An experiment in authorship attribution. In *JADT 2002 : 6es Journées internationales d'Analyse statistique des Données Textuelles*, Saint Malo, pp. 29–37. JADT. 12, 13, 14

Baayen, H., H. Van Halteren, and F. Tweedie (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing 11*(3), 121–132. 8, 13

Biber, D. (1990). Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. *Literary and linguistic computing 5*(4), 257–269. 6, 28

Biber, D. (1993a). Co-occurrence Patterns among Collocations: a Tool for Corpus-based Lexical Knowledge Acquisition. *Computational Linguistics 19*(3), 531–538. 6, 12

Biber, D. (1993b). Representativeness in Corpus Design. *Literary and linguistic computing 8*(4), 243–257. 28

Bird, S., E. Klein, and E. Loper (2009). *Natural language processing with Python*. O'Reilly Media, Inc. 18

Björnsson, C. H. (1968). *Läsbarhet*. Liber. 26, 27

Bouckaert, R. R. (2003). Choosing between Two Learning Algorithms Based on Calibrated Tests. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, Washington DC, pp. 51–58. Morgan Kaufman. 18

Brennan, M., S. Afroz, and R. Greenstadt (2012, November). Adversarial stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Transactions on Information and System Security (TISSEC) 15*(3), 12:1–12:22. 4, 5, 20

Brocardo, M. L., I. Traore, and I. Woungang (2014). Toward a Framework for Continuous Authentication Using Stylometry. In *Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on*, Victoria, BC, Canada, pp. 106–115. IEEE: IEEE. 20

Burrows, J. (2002). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing 17*(3), 267–287. 35

Burrows, J. F. (1987). *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Oxford: Oxford University Press. 13

Burrows, J. F. (1992, April). Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information. *Literary and Linguistic Computing 7*(2), 91–109. 13

Calix, K., M. Connors, D. Levy, H. Manzar, G. MCabe, and S. Westcott (2008, May). Stylometry for E-mail Author Identification and Authentication. In *Proceedings of CSIS Research Day*. Pace University. 10

Chaski, C. E. (2005). Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence 4*(1), 1–13. 8

Corney, M., O. de Vel, A. Anderson, and G. Mohay (2002). Gender-Preferential Text Mining of E-mail Discourse. In *Computer Security Applications Conference, 2002. Proceedings. 18th Annual*, pp. 282–289. IEEE. 6

Dabagh, R. M. (2007). Authorship Attribution and Statistical Text Analysis. *Metodološki zvezki 4*(2), 149–163. 10

Daelemans, W. (2013). Explanation in Computational Stylometry. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, Volume 7817, pp. 451–462. Berlin Heidelberg: Springer. 4

Daelemans, W. and V. Hoste (2013, July). STYLENE: an Environment for Stylometry and Readability Research for Dutch. Technical report, CLiPS Research Center, University of Antwerp, Antwerp, Belgium. 19

Daelemans, W. and A. Van den Bosch (2005). *Memory-Based Language Processing*. Studies in Natural Language Processing. Cambridge: Cambridge University Press. 64

Daelemans, W., J. Zavrel, K. van der Sloot, and A. Van den Bosch (2003, May). TiMBL: Tilburg Memory-Based Learner - version 6.3. Ilk technical report – ilk 10-01, Tilburg University and CLiPS, University of Antwerp, Tilburg, The Netherlands. 19

De Morgan, S. E. (1882). *Memoir of Augustus De Morgan*. London: Longmans, Green, and Company. 5

De Smedt, K. and V. Rosén (1999). Automatic proofreading for Norwegian: The challenges of lexical and grammatical variation. In T. Nordgård (Ed.), *NODALIDA '99: Proceedings from the 12th "Nordiske datalingvistikkdager"*, Trondheim, pp. 206–215. NTNU. 10

De Vel, O., A. Anderson, M. Corney, and G. Mohay (2001, December). Mining e-Mail Content for Author Identification Forensics. *Sigmod Record 30*(4), 55–64. 9

Dumais, S., J. Platt, D. Heckerman, and M. Sahami (1998). Inductive Learning Algorithms and Representations for Text Categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, CIKM '98, New York, NY, USA, pp. 148–155. ACM: ACM. 17

Dyvik, H. (2000). Nødvendige noder i norsk: Grunntrekk i en leksikalsk-funksjonell beskrivelse av norsk syntaks [Necessary nodes in Norwegian: Basic properties of a lexical-functional description of Norwegian syntax]. In Ø. Andersen, K. Fløttum, and T. Kinn (Eds.), *Menneske, språk og felleskap*, pp. 25–45. Oslo: Novus forlag. 24

Dyvik, H. (2012). Norm clusters in written Norwegian. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian 49*, 193–220. 10

Eder, M., M. Kestemont, and J. Rybicki (2013). Stylometry with R: a suite of tools. In *Digital Humanities 2013: Conference Abstracts*, pp. 487–489. 1, 20, 22, 35, 36, 42

Eder, M., J. Rybicki, and M. Kestemont (2014, August). *'Stylo': a package for stylometric analyses*. Computational Stylistic Group. 35

Eklund, P. W. and A. Hoang (2002). A Performance Survey of Public Domain Supervised Machine Learning Algorithms. *Australian Journal of Intelligent Information Systems. 9*(1), 1–47. 17

El-Fiqi, H., E. Petraki, and H. A. Abbass (2011, June). A Computational Linguistic Approach for the Identification of Translator Stylometry in Arabic-English Text. In *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, pp. 2039–2045. IEEE. 20

Faarlund, J. T., S. Lie, and K. I. Vannebo (1997). *Norsk referansegrammatikk*. Oslo: Universitetsforlaget. 53

Flesch, R. (1948, June). A new readability yardstick. *Journal of Applied Psychology 32*(3), 221–233. 7

Gamon, M. (2004). Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, pp. 841—847. Association for Computational Linguistics: International Conference on Computational Linguistics. 8

Goswami, S., S. Sarkar, and M. Rustagi (2009). Stylometric Analysis of Bloggers' Age and Gender. In *Proceedings of the Third International ICWSM Conference*. AAAI Press. 6

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD explorations newsletter 11*(1), 10–18. 1, 20

Hirst, G. and O. Feiguina (2007, October). Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing 22*(4), 405–417. 8, 16

Hirst, G. and V. Wei Feng (2012). Changes in style in authors with Alzheimer's disease. *English Studies 93*(3), 357–370. 5

Hofland, K. (2000). Self-Expanding Corpus Based on Newspapers on the Web. In *Proceedings of the Second International Language Resources and Evaluation Conference.*, Paris. European Language Resources Association. 23

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities 28*(2), 87–106. 10, 12

Holmes, D. I. (1998, September). The Evolution of Stylometry in Humanities Scholarship. *Literary and linguistic computing 13*(3), 111–117. 17

Holmes, D. I. and F. J. Tweedie (1995). Forensic Stylometry: A Review of the Cusum Controversy. *Revue Informatique et Statistique dans les Sciences Humaines 31*(1), 19–47. 21

Hoover, D. L. (2001). Statistical Stylistics and Authorship Attribution: an Empirical Investigation. *Literary and Linguistic Computing 16*(4), 421–444. 13

Hoover, D. L. (2004a). Delta Prime? *Literary and Linguistic Computing 19*(4), 477–495. 35

Hoover, D. L. (2004b). Testing Burrows's Delta. *Literary and linguistic computing 19*(4), 453–475. 35

Jockers, M. L., D. M. Witten, and C. S. Criddle (2008). Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing 23*(4), 465–491. 35

Juola, P. (2013, July). Rowling and "Galbraith": an authorial analysis. Lest 11. mai 2015. <http://languagelog.ldc.upenn.edu/nll/?p=5315>. 20

Juola, P. et al. (2009). JGAAP: A System for Comparative Evaluation of Authorship Attribution. In *JDHCS 2009*, Volume 1, pp. 1–5. The Division of the Humanities at the University of Chicago. 22

Keshtkar, F. and D. Inkpen (2009, September). Using Sentiment Orientation Features for Mood Classification in Blogs. In *Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on*, pp. 1–6. 2

Khmelev, D. V. and W. J. Teahan (2003). A Repetition Based Measure for Verification of Text Collections and for Text Categorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 104–110. ACM. 7

Kincaid, J. P., R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom (1975, February). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, DTIC Document, Millington, TN, US. 7

Koppel, M., S. Argamon, and A. R. Shimoni (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing 17*(4), 401–412. 6, 18

Koppel, M., J. Schler, and S. Argamon (2009, January). Computational Methods in Authorship Attribution. *Journal of the American Society for information Science and Technology 60*(1), 9–26. 8

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, Amsterdam, The Netherlands, pp. 3–24. IOS Press. 16, 17, 18, 19, 58, 67

Li, J., R. Zheng, and H. Chen (2006). From Fingerprint to Writeprint. *Communications of the ACM 49*(4), 76–82. 8, 9

Lim, T.-S., W.-Y. Loh, and Y.-S. Shih (2000, September). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. *Machine Learning 40*(3), 203–229. 17

Linmans, A. J. M. (1998). Correspondence Analysis of the Synoptic Gospels. *Literary and linguistic computing 13*(1), 1–13. 11

López-Escobedo, F., C.-F. Méndez-Cruz, G. Sierra, and J. Solórzano-Soto (2013, October). Analysis of Stylometric Variables in Long and Short Texts. *Procedia-Social and Behavioral Sciences 95*, 604–611. 11, 14

Lowe, D. and R. Matthews (1995). Shakespeare Vs. Fletcher: A Stylometric Analysis by Radial Basis Functions. *Computers and the Humanities 29*(6), 449–461. 17

Luyckx, K. and W. Daelemans (2008a, August). Authorship Attribution and Verification with Many Authors and Limited Data. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Volume 1, Stroudsburg, PA, US, pp. 513–520. Association for Computational Linguistics. 19

Luyckx, K. and W. Daelemans (2008b). Personae: a Corpus for Author and Personality Prediction from Text. In *International Conference on Language Resources and Evaluation*. European Language Resources Association. 2, 6, 19, 22

Lyall, S. (2013, July). This Detective Novel's Story Doesn't Add Up. Lest 11. mai 2015. <http://www.nytimes.com/2013/07/15/books/a-detective-storys-famous-author-is-unmasked.html?pagewanted=1&_r=1>. 1

Mascol, C. (1888). Curves of Pauline and Pseudo-Pauline Style i. *Unitarian Review 30*, 453–460. 5

Matsuura, T. and Y. Kanada (2000). Extraction of Authors' Characteristics from Japanese Modern Sentences via N-gram Distribution. In *Discovery Science*, pp. 315–319. Springer. 7

Matthews, R. A. and T. V. Merriam (1993). Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher. *Literary and Linguistic Computing 8*(4), 203–209. 16, 17

McCallum, A., K. Nigam, et al. (1998). A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI-98 workshop on learning for text categorization*, Volume 752, pp. 41–48. Citeseer. 62

McDonald, A. W., J. Ulman, M. Barrowclift, and R. Greenstadt (2013). Anonymouth Revamped: Getting Closer to Stylometric Anonymity. In *PETools: Workshop on Privacy Enhancing Tools*. 20

McMenamin, G. R. (2002). *Forensic linguistics: Advances in forensic stylistics*. CRC press. 4

Mealand, D. L. (1999). Style, Genre, and Authorship in Acts, the Septuagint, and Hellenistic Historians. *Literary and linguistic computing 14*(4), 479–506. 11

Mendenhall, T. C. (1887, March). The Characteristic Curves of Composition. *Science 9*(214), 237–246. 5, 7, 10

Merriam, T. V. and R. A. Matthews (1994). Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe. *Literary and Linguistic Computing 9*(1), 1–6. 16

Meurer, P. (2012). Corpuscle–a new corpus management platform for annotated corpora. In G. Andersen (Ed.), *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian*, Volume 49 of *Studies in Corpus Linguistics*, pp. 29–50. John Benjamins Publishing. 23

Meurer, P., M. Butt, and T. H. King (2012). INESS-Search: A search system for LFG (and other) treebanks. In *Proceedings of the LFG'12 Conference*, pp. 404–421. 23

Mosteller, F. and D. Wallace (1964). *Inference and disputed authorship: The Federalist*. Addison-Wesley. 5, 8

Nerbonne, J. (2007). *The Exact Analysis of Text*. Stanford, CA, US: CSLI. 3

Nguyen, D., R. Gravel, D. Trieschnigg, and T. Meder (2013). " How Old Do You Think I Am?"; A Study of Language and Age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. AAAI Press. 4, 6

Norusis, M. (2008). *SPSS 16.0 statistical procedures companion*. Prentice Hall Press. 12

Oakes, M. P. (2014). *Literary Detective Work on the Computer*, Volume 12. John Benjamins Publishing Company. 6

Peersman, C., W. Daelemans, and L. Van Vaerenbergh (2011). Predicting Age and Gender in Online Social Networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pp. 37–44. ACM. 6

Pennebaker, J. W. and L. D. Stone (2003). Words of Wisdom: Language Use Over the Life Span. *Journal of personality and social psychology 85*(2), 291. 4

Poulsen, J. and A. French (2008). Discriminant Function Analysis. Lest 11. mai 2015. <http://faculty.ksu.edu.sa/hisham/Documents/Students%20Work%202/discrim.pdf>. 12

Rokach, L. and O. Maimon (2005). *Decision Trees*. Springer. 18

Rosenblatt, F. (1961). Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms. Technical report, DTIC Document. 16

Rudman, J. (1998). The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities 31*(4), 351–365. 21, 22, 73

Rudman, J. (2010). The State of Non-traditional Authorship Studies—2010: Some Problems and Solutions. *Proceedings of the Digital Humanities*, 217–219. 21, 22

Rudman, J. (2012). The state of Non-Traditional Authorship Attribution Studies—2012: Some Problems and Solutions. *English Studies 93*(3), 259–274. 21

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1985). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1, Chapter Learning Internal Representations by Error Propagation, pp. 318–362. MIT Press. 16

Schler, J., M. Koppel, S. Argamon, and J. W. Pennebaker (2006). Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, Volume 6, pp. 199–205. 2

Sebastiani, F. (2005). Text categorization. In *Text Mining and its Applications to Intelligende, CRM and Knowledge Management*, pp. 109–129. WIT Press. 10, 15

Sherman, L. A. (1888). *Some observations upon the sentence-length in English prose*. Lincoln. 5, 7

Smith, L. I. (2002). A tutorial on Principal Components Analysis. *Cornell University, USA*, 1–26. 13

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for information Science and Technology 60*(3), 538–556. 6, 7, 8, 9, 16, 21

Stamatatos, E., N. Fakotakis, and G. Kokkinakis (2000). Automatic Text Categorization in Terms of Genre and Author. *Computational linguistics 26*(4), 471–495. 8

Stamatatos, E., N. Fakotakis, and G. Kokkinakis (2001). Computer-based Authorship Attribution without Lexical Measures. *Computers and the Humanities 35*(2), 193–214. 8

StatSoft (2013). *Electronic Statistics Textbook*. Tulsa, OK: StatSoft, Inc. 14

Tabata, T. (2007). A Statistical Study of Superlatives in Dickens and Smollett: A Case Study in Corpus Stylistics. In *Digital Humanities 2007*, Urbana-Champaign, pp. 210–215. The Association for Computers and the Humanities and The Association for Literary and Linguistic Computing: University of Illinois. 11

Tambouratzis, G., S. Markantonatou, N. Hairetakis, M. Vassiliou, G. Carayannis, and D. Tambouratzis (2004). Discriminating the Registers and Styles in the Modern Greek Language-Part 1: Diglossia in Stylistic Analysis. *Literary and Linguistic Computing 19*(2), 197–220. 9

Teng, G.-F., M.-S. Lai, J.-B. Ma, and Y. Li (2004). E-mail authorship mining based on SVM for computer forensic. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, Volume 2, pp. 1204–1207. IEEE. 9

Todorov, T. (1971). The Place of Style in the Structure of the Text. In *Literary Style: A Symposium*, Volume 32. 21

Torres-Reyna, O. (2010). Getting Started in Factor Analysis (using Stata 10). Lest 11. mai 2015. `<http://dss.princeton.edu/training/Factor.pdf>`. 12

Uzuner, Ö., B. Katz, and T. Nahnsen (2005). Using Syntactic Information to Identify Plagiarism. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pp. 37–44. Association for Computational Linguistics. 8

Van Halteren, H. (2004). Linguistic Profiling for Author Recognition and Verification. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 199. Association for Computational Linguistics. 8

Van Halteren, H., H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt (2005). New Machine Learning Methods Demonstrate the Existence of a Human Stylome. *Journal of Quantitative Linguistics 12*(1), 65–77. 3

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. Springer Science & Business Media. 15

Yang, Y. and G. I. Webb (2003). On Why Discretization Works for Naive-Bayes Classifiers. In T. D. Gedeon and F. L. C. C (Eds.), *AI 2003: Advances in Artificial Intelligence*, Volume 2903 of *Lecture Notes in Computer Science*, pp. 440–452. Berlin Heidelberg: Springer. 18

Young, F. W. and C. M. Bann (1996). ViSta: The Visual Statistics System. Technical report, Technical Report 94–1 (c). 11

Zheng, R., J. Li, H. Chen, and Z. Huang (2006, February). A Framework for Authorship Identification of Online Messages: Writing-style Features and Classification Techniques. *Journal of the American Society for Information Science and Technology 57*(3), 378–393. 5, 6, 9, 10, 15

Tillegg A

Korpusoversikt

	Forfatter	Tittel	Sjanger
ar1	Ragde, Anne B.	En kald dag i helvete	Roman
ar2	Ragde, Anne B.	Ansiktet som solen	Kortprosa
mj1	Johnsgaard, Magnar	Vanskapningen	Roman
mj2	Johnsgaard, Magnar	Veivokteren	Kortprosa
ph1	Hagan, Patricia	Vill og vakker	Roman
ph2	Hagan, Patricia	Kjærlighetens triumf	Roman
ul1	Lindell, Unni	Slangebæreren	Roman
ul2	Lindell, Unni	En grusom kvinnes bekjennelser	Kortprosa
ao1	Oterholm, Anne	Avbrutt selskap	Roman
ao2	Oterholm, Anne	Avslutningen	Roman
ok1	Klippenvåg, Odd	Body and Soul	Kortprosa
ok2	Klippenvåg, Odd	Bruckner, en lengsel	Roman
ok3	Klippenvåg, Odd	Et virkelig liv	Roman
el1	Lindvåg, Ellen Iris	Ingen kan nå med	Roman
el2	Lindvåg, Ellen Iris	Sett: ?	Roman
re1	Enger, Rolf	Solformørkelse	Roman
re2	Enger, Rolf	Hvis noen skulle være så vemmelige	Kortprosa
dh1	Hamilton, Dan	Kameleonkvinnen	Roman
dh2	Hamilton, Dan	Tiggerkongen	Roman
ek1	Kiøsterud, Erland	Ved fjellets fot	Kortprosa
ek2	Kiøsterud, Erland	Den norske sangeren	Kortprosa

Tabell A.1: Oversikt over innhold i korpuset

Tillegg B

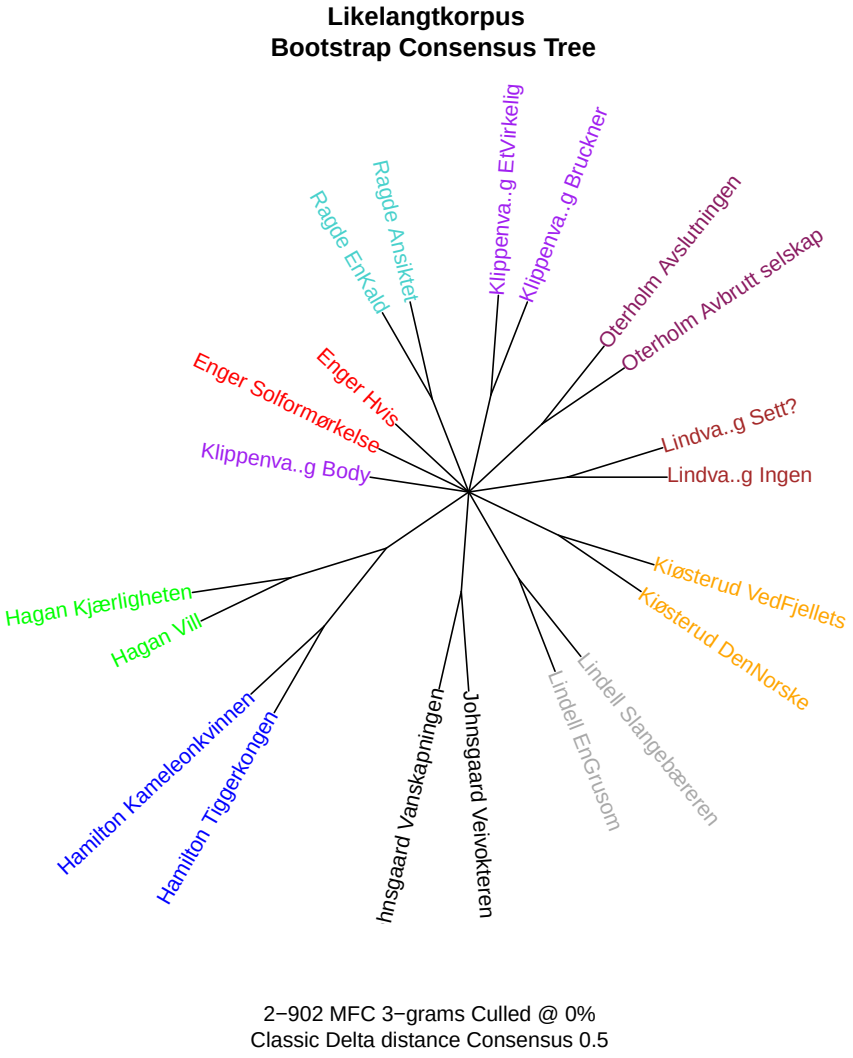
Programmet “wordsplitter.sh”

```
#!/bin/bash
#Author: Victoria Troland
#Vår 2015
# Remove files
rm -f splitab splitac splitad splitae splitaf splitag splitai splitaj
splitak new splital

# Replace space with newline
tr " " "\n"  > new

# Cut after linje 4500 with slit -l
split -l 4500 new split

# Replace newline with space
tr "\n" " " < splitaa > newtext

echo "Ferdig!"
```

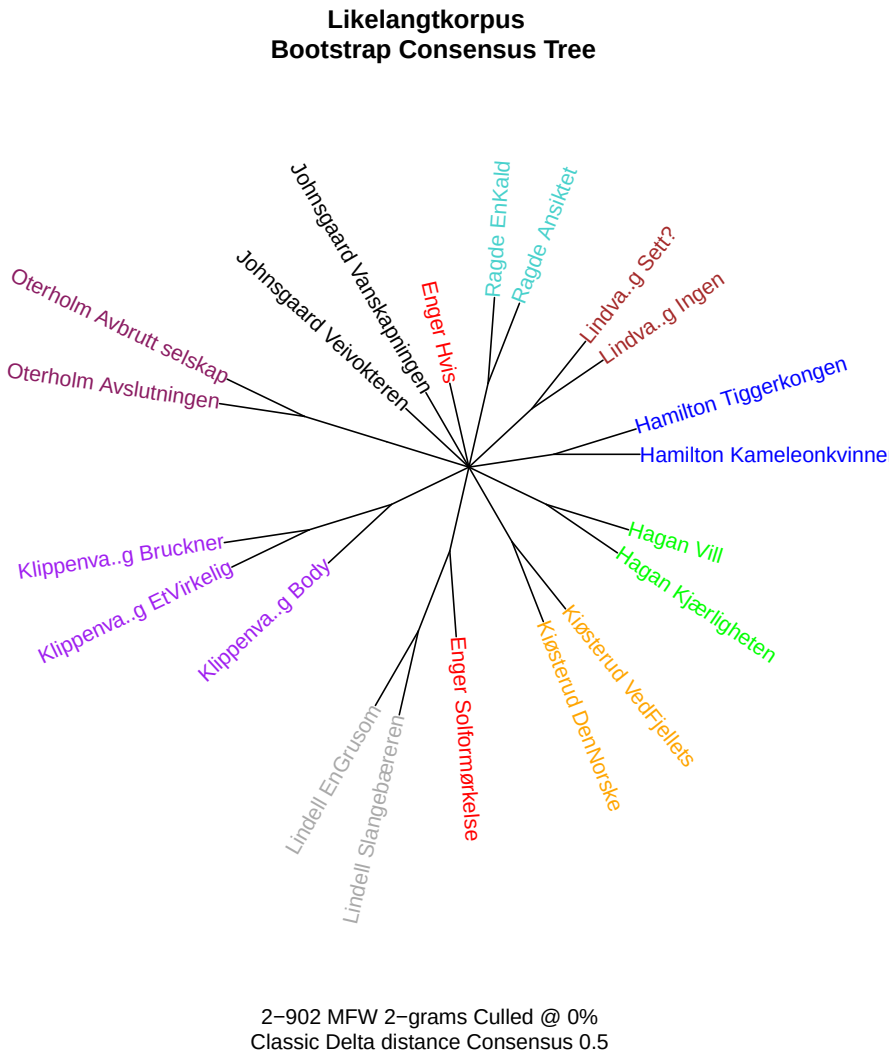
Tillegg C

Grafer fra “Likelangtkorpuset”

C.1 “Bootstrap Consensus tree”-forsøk

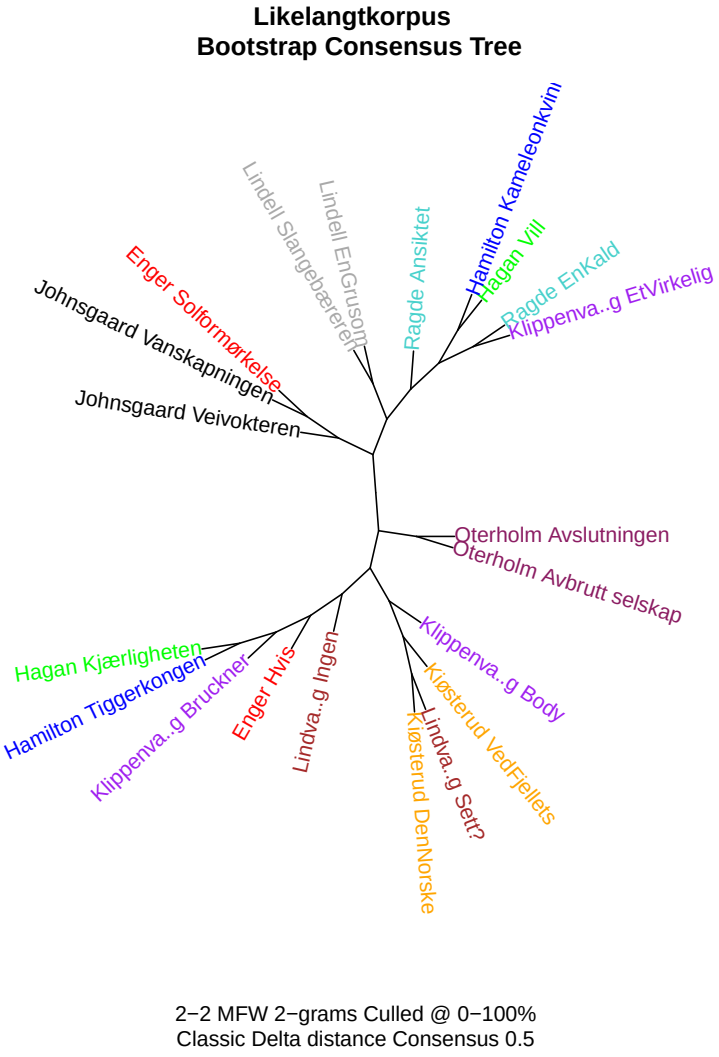


Figur C.1: Med n-gram (n=3) av bokstaver MFW=0-1000

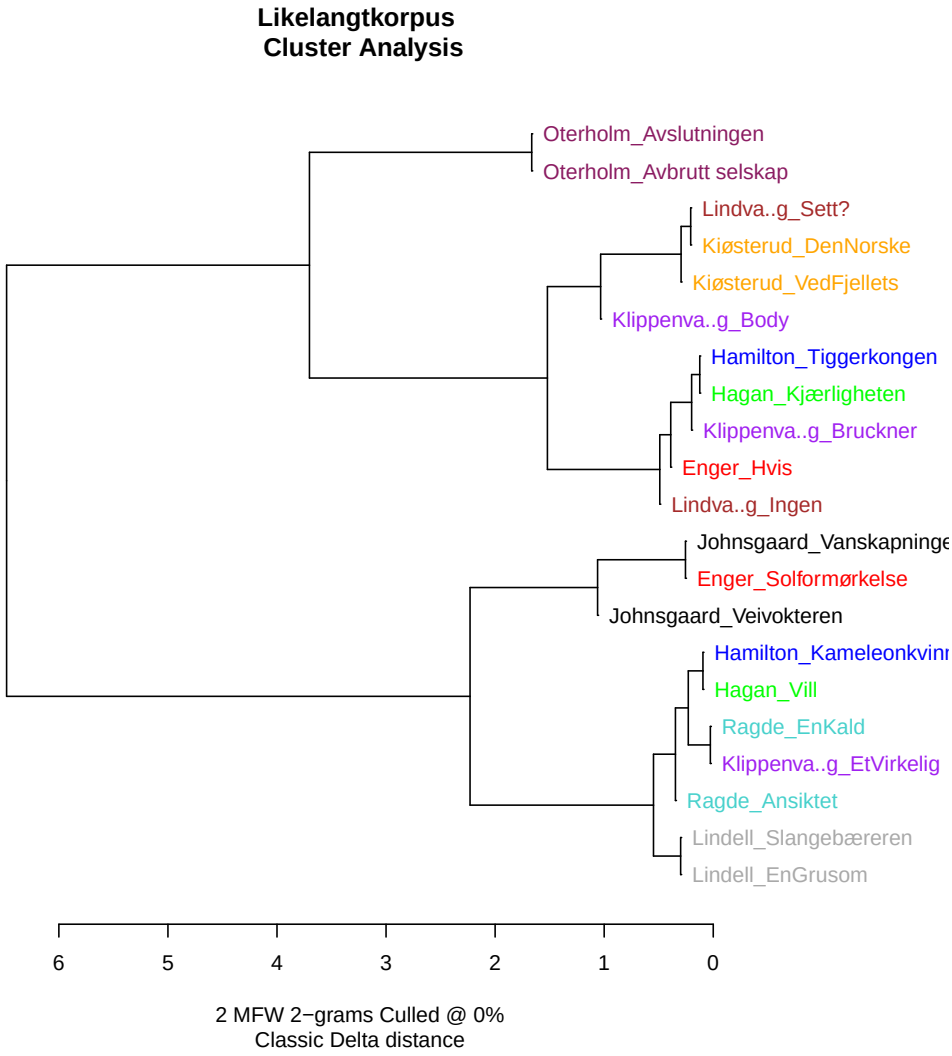


Figur C.2: Med n-gram (n=2) av ord MFW=0-1000

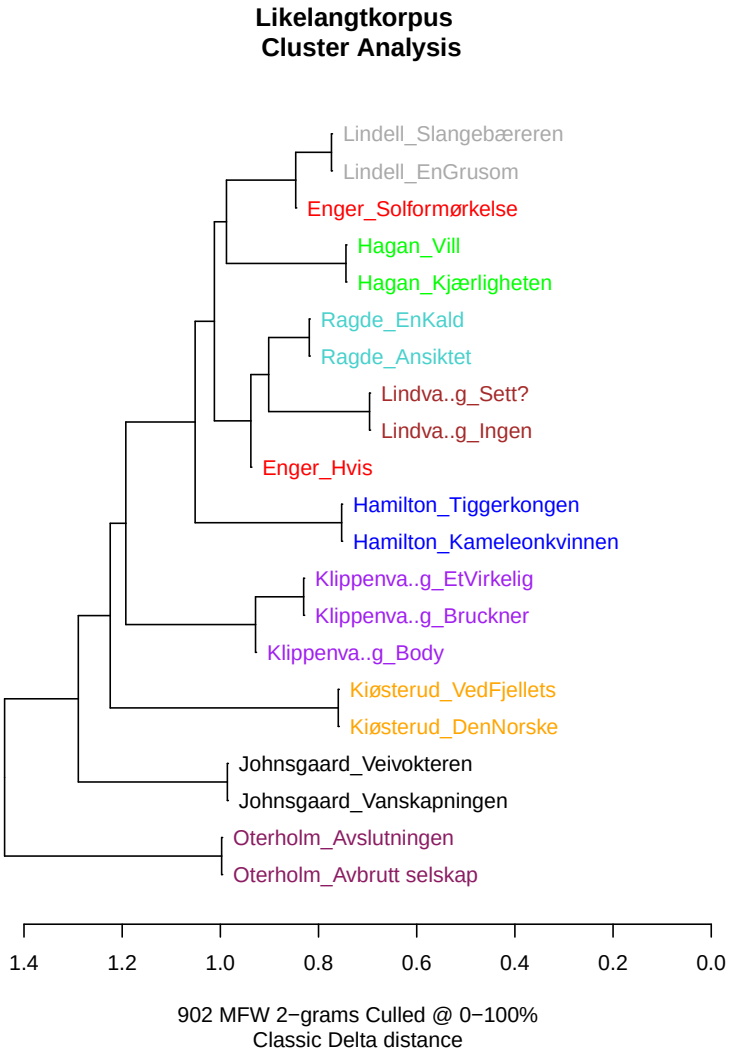
C.2 MFW og "culling" parameterforsøk



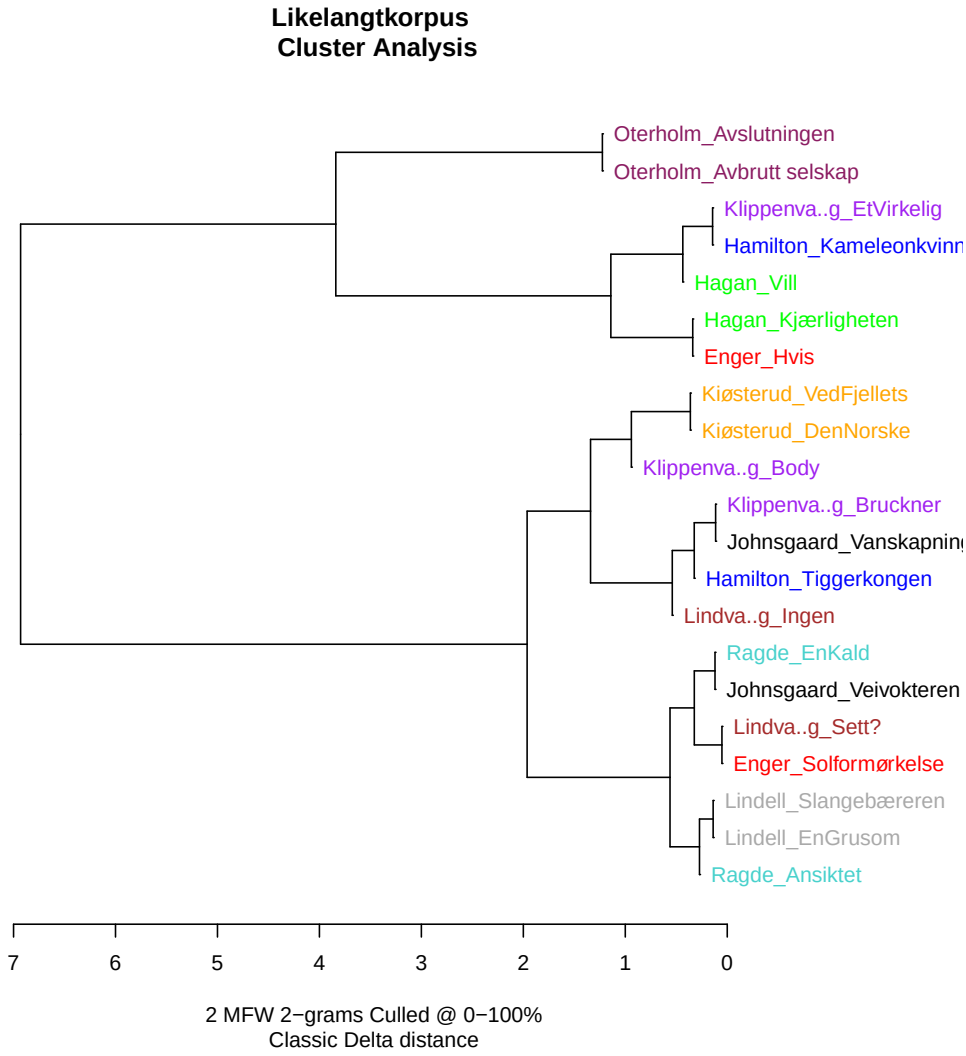
Figur C.3: Med n-gram (n=2) av ord MFW=0, C=0-100



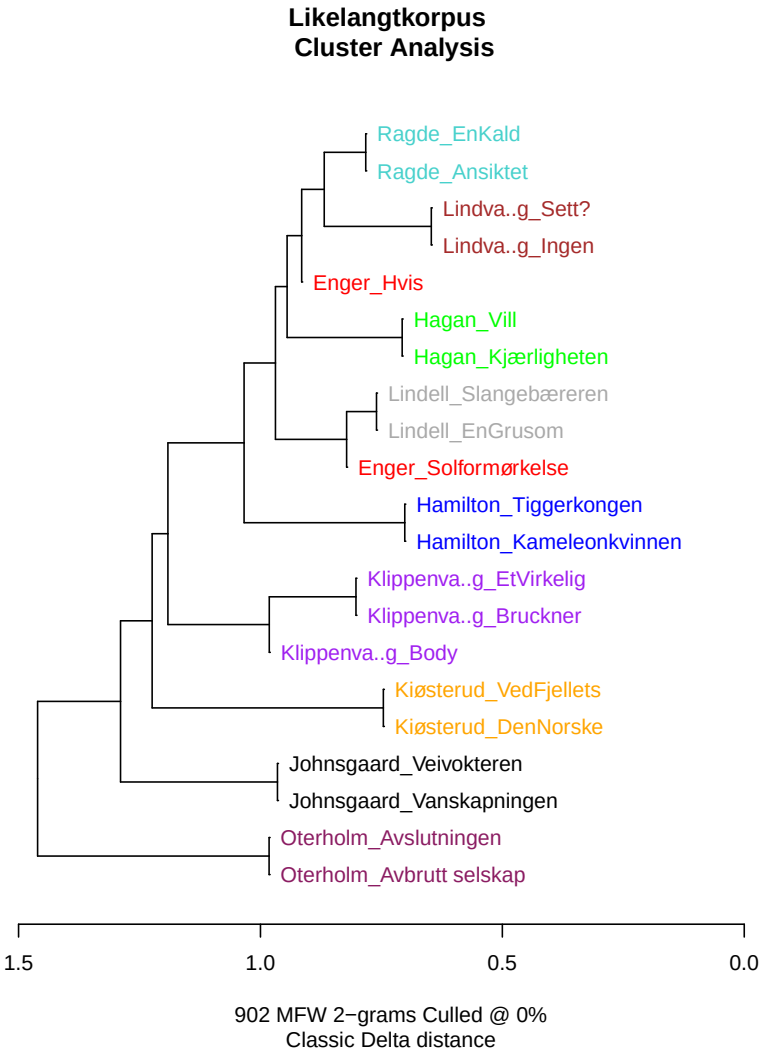
Figur C.4: Med n-gram (n=2) av ord MFW=0, C=0



Figur C.5: Med n-gram (n=2) av ord MFW=0-1000, C=0-100



Figur C.6: Med n-gram (n=2) av ord MFW=0, C=0-100

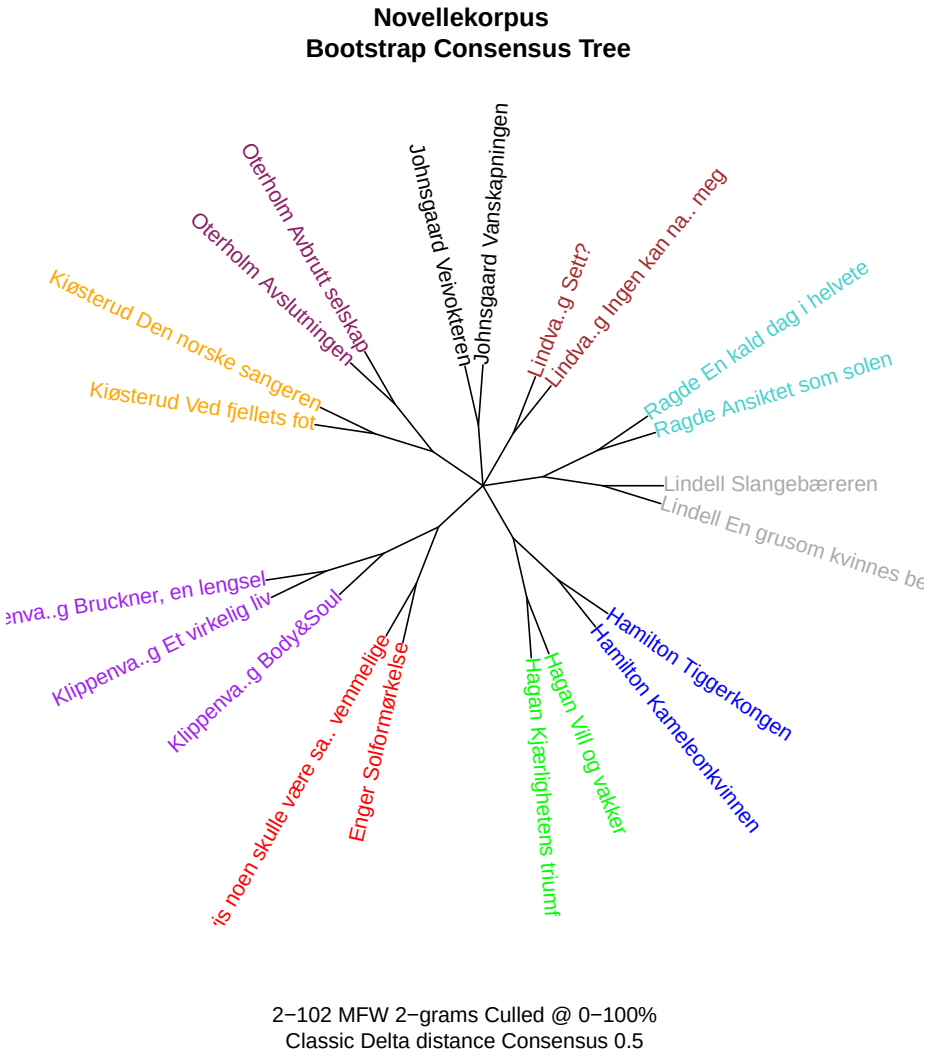


Figur C.7: Med n-gram (n=2) av ord MFW=0-1000, C=0

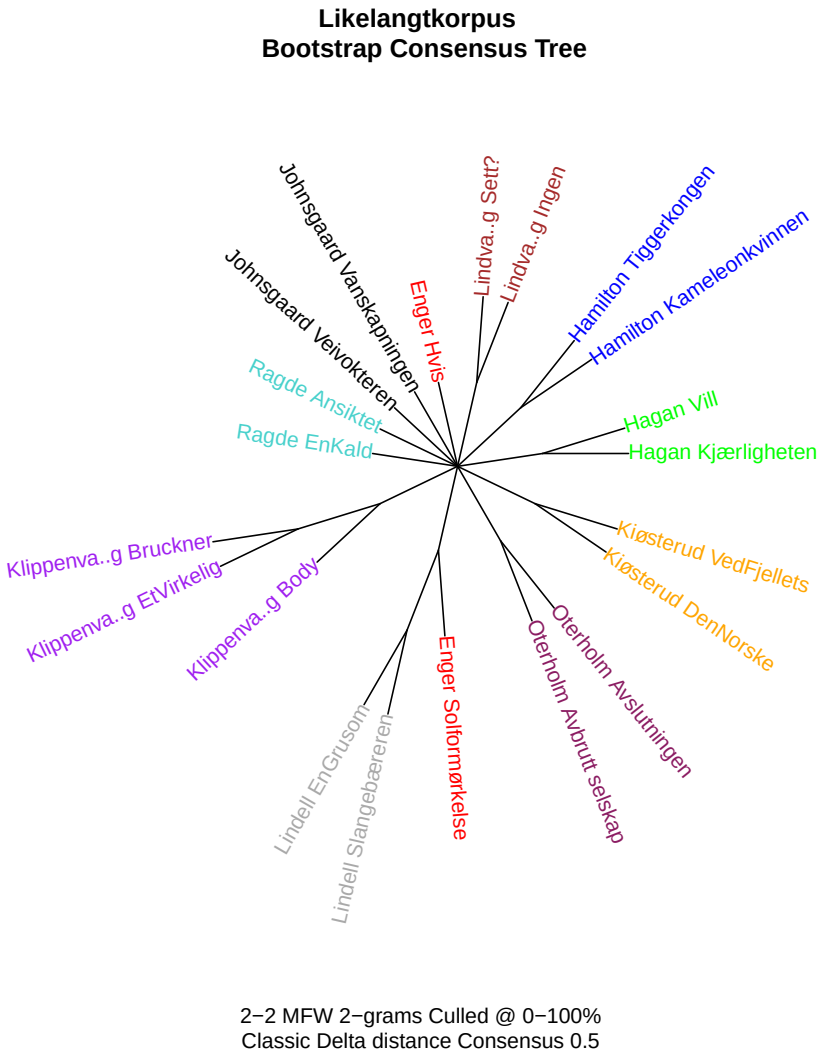
Tillegg D

Grafer fra “Novellekorpuset”-forsøk

D.1 “Novellekorpus” sammenlignet med “Likelangtkorpuset”

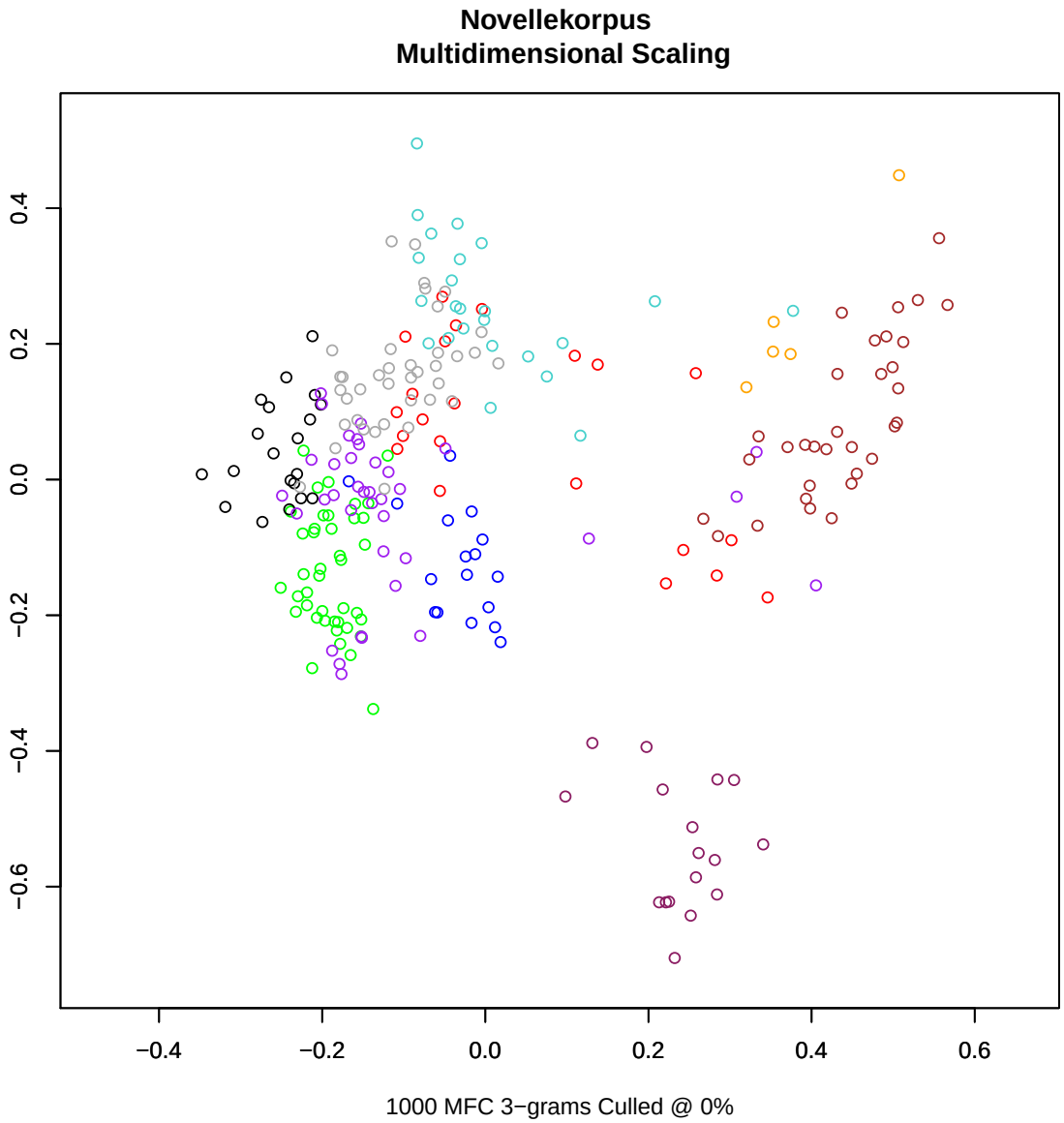


Figur D.1: “Novellekorpuset” med n-gram (n=2) av ord MFW=0-1000, C=0-100



Figur D.2: Fra "Likelangtkorpuset" med n-gram (n=2) av ord MFW=0-1000, C=0-100

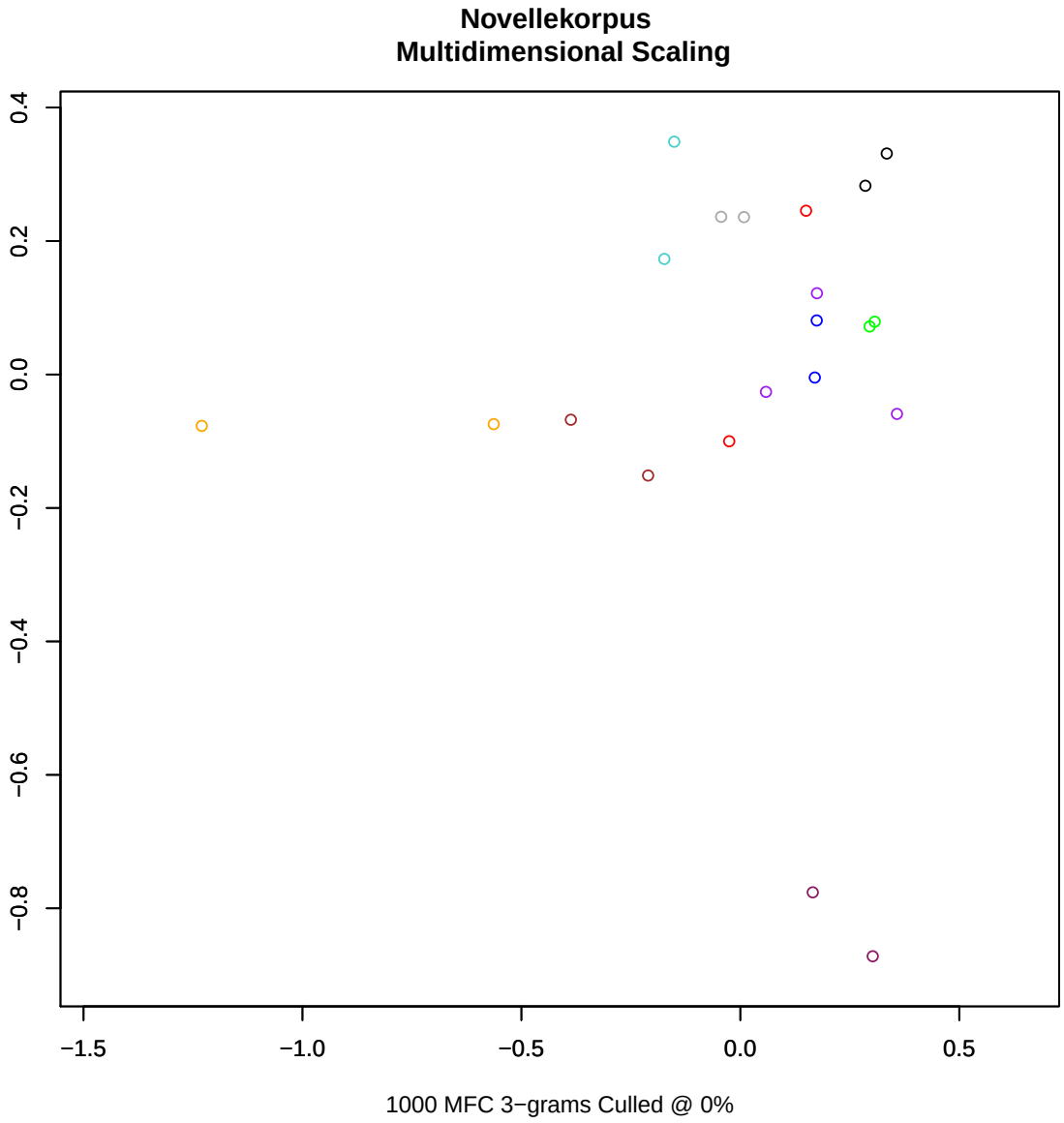
D.2 "Sampling" forsøk



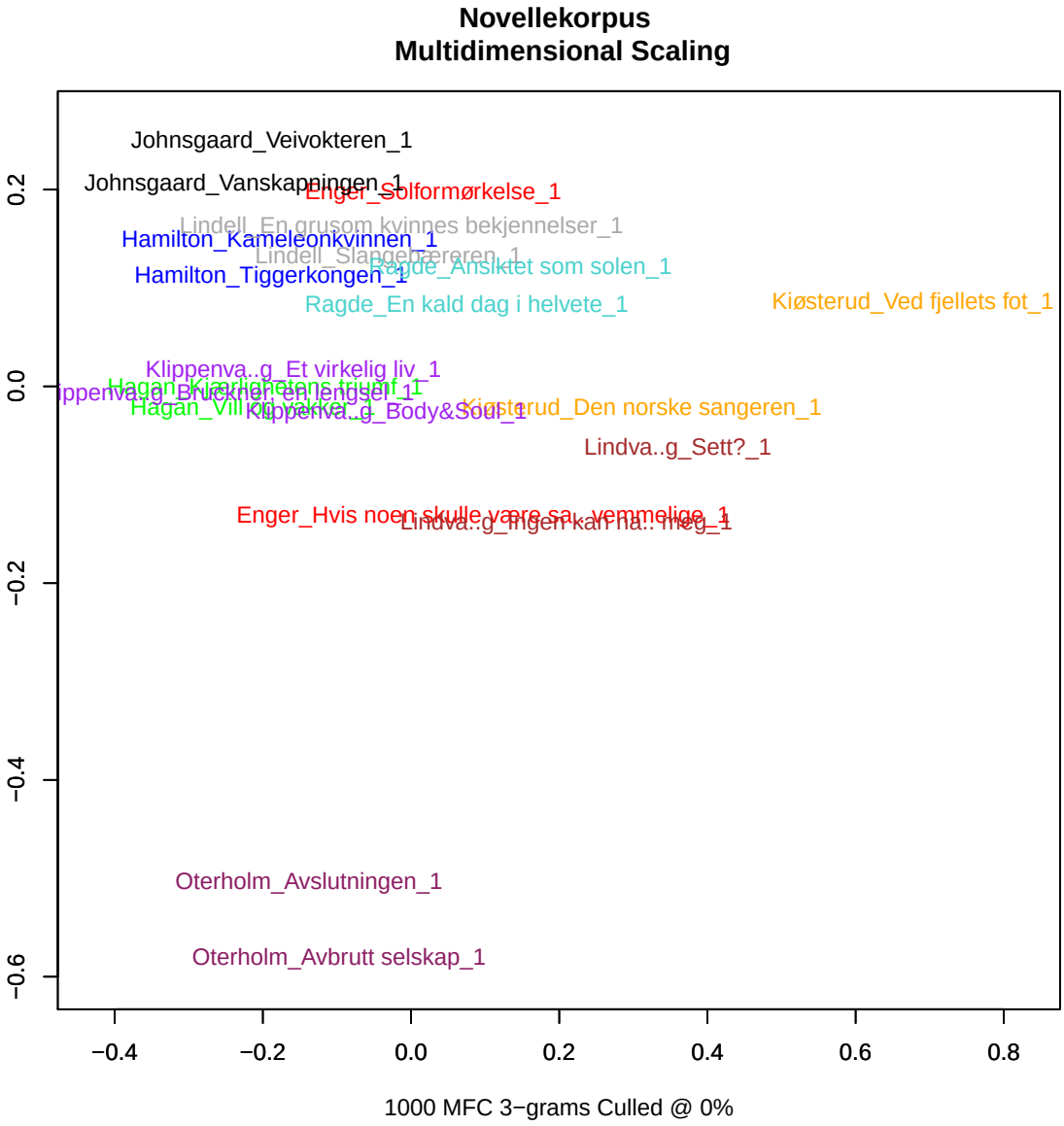
Figur D.3: Fra "Novellekorpus" med n-gram (n=3) av bokstaver MFW=0-1000, C=0, med "normal sampling"

Tabell D.1: Fargekodene til forfatterene

	Forfatter	Farge
1	Oterholm	Mørkelilla
2	Johnsgaard	Svart
3	Klippenvåg	Lilla
4	Enger	Rød
5	Ragde	Lyseblå
6	Lindvåg	Mørkerød
7	Hamilton	Blå
8	Hagan	Grønn
9	Kiøsterud	Oransje
10	Lindell	Grå



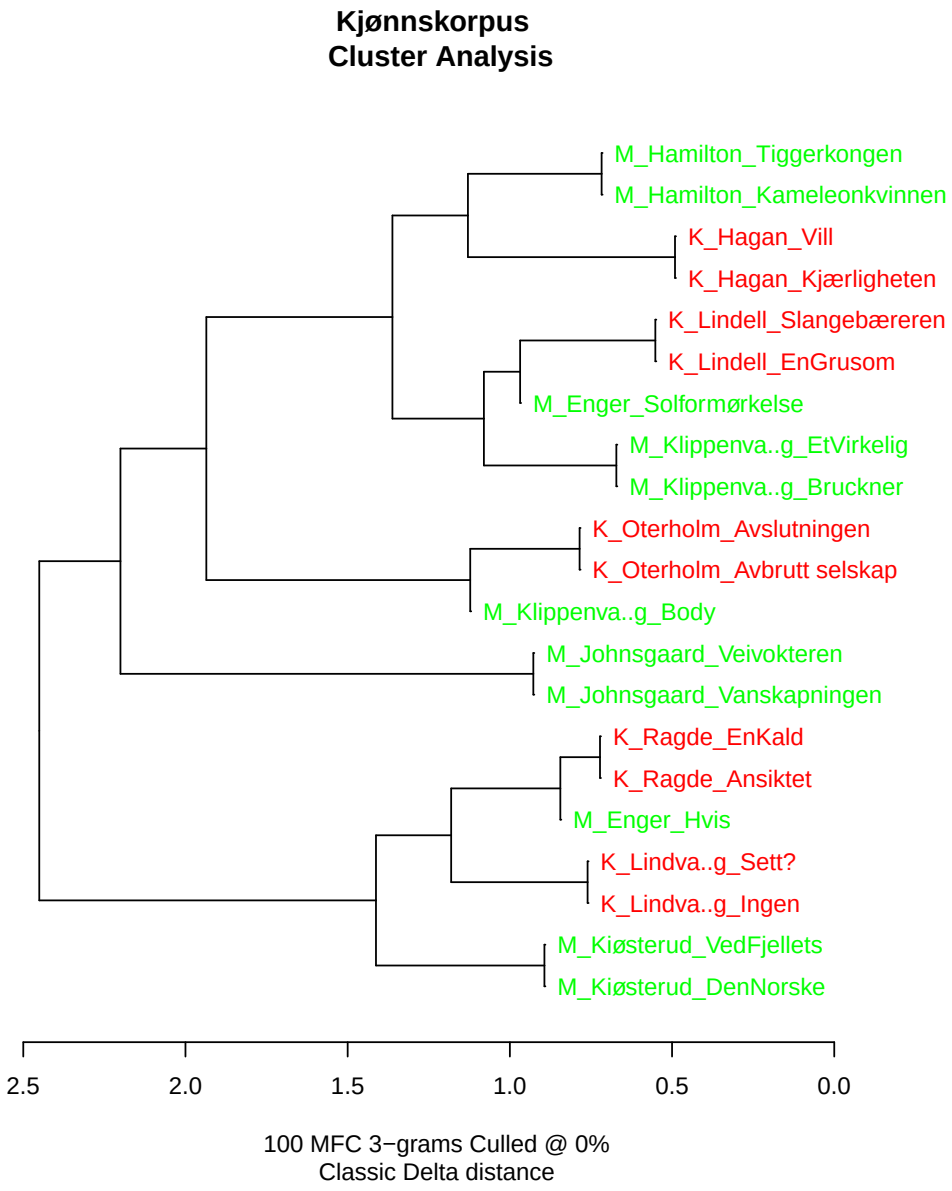
Figur D.4: Fra "Novellekorpus" med n-gram (n=3) av bokstaver MFW=0-1000, C=0 uten sampling



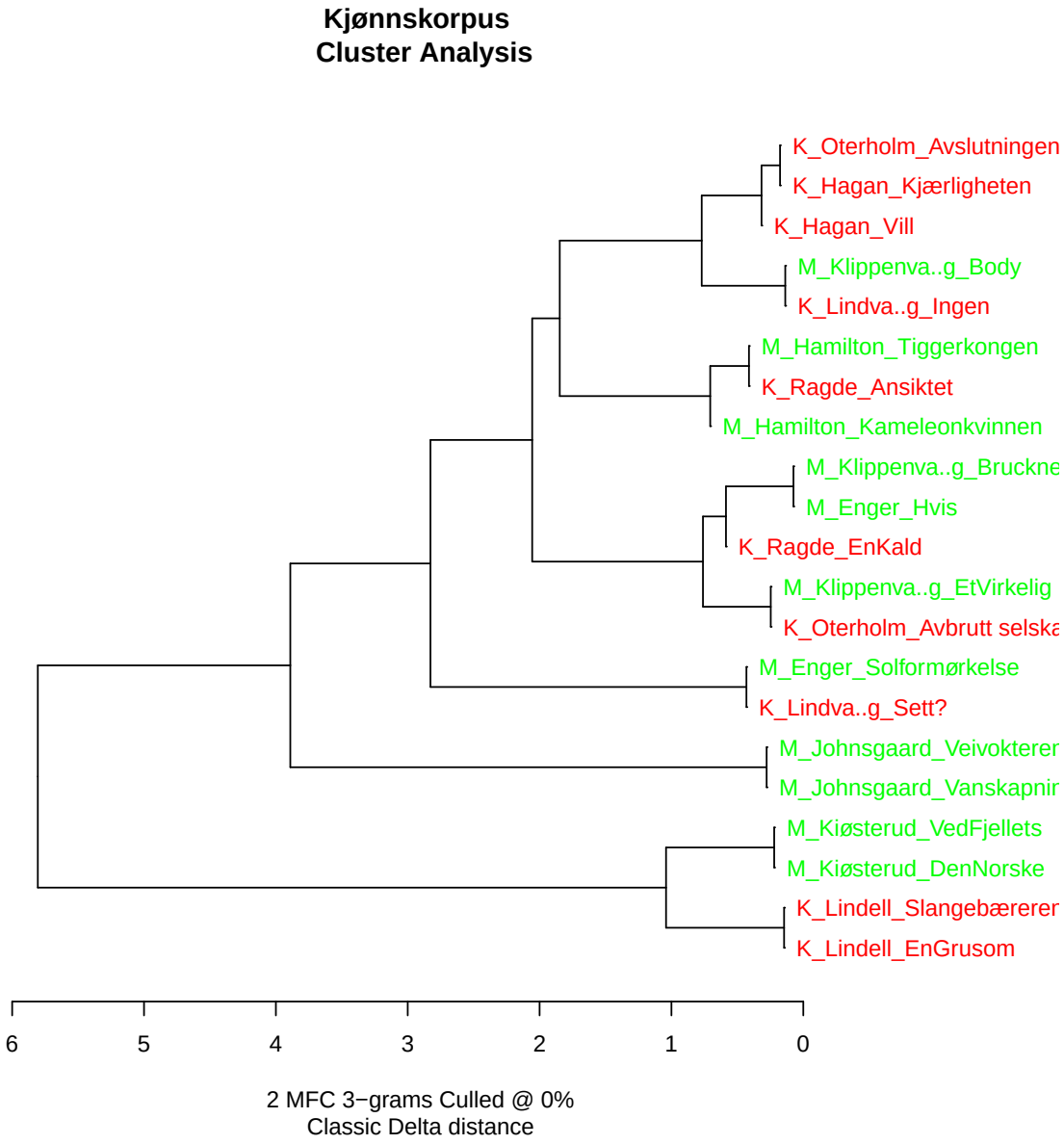
Figur D.5: Fra "Novellekorpus" med n-gram (n=3) av bokstaver MFW=0-1000, C=0, "random sampling"

Tillegg E

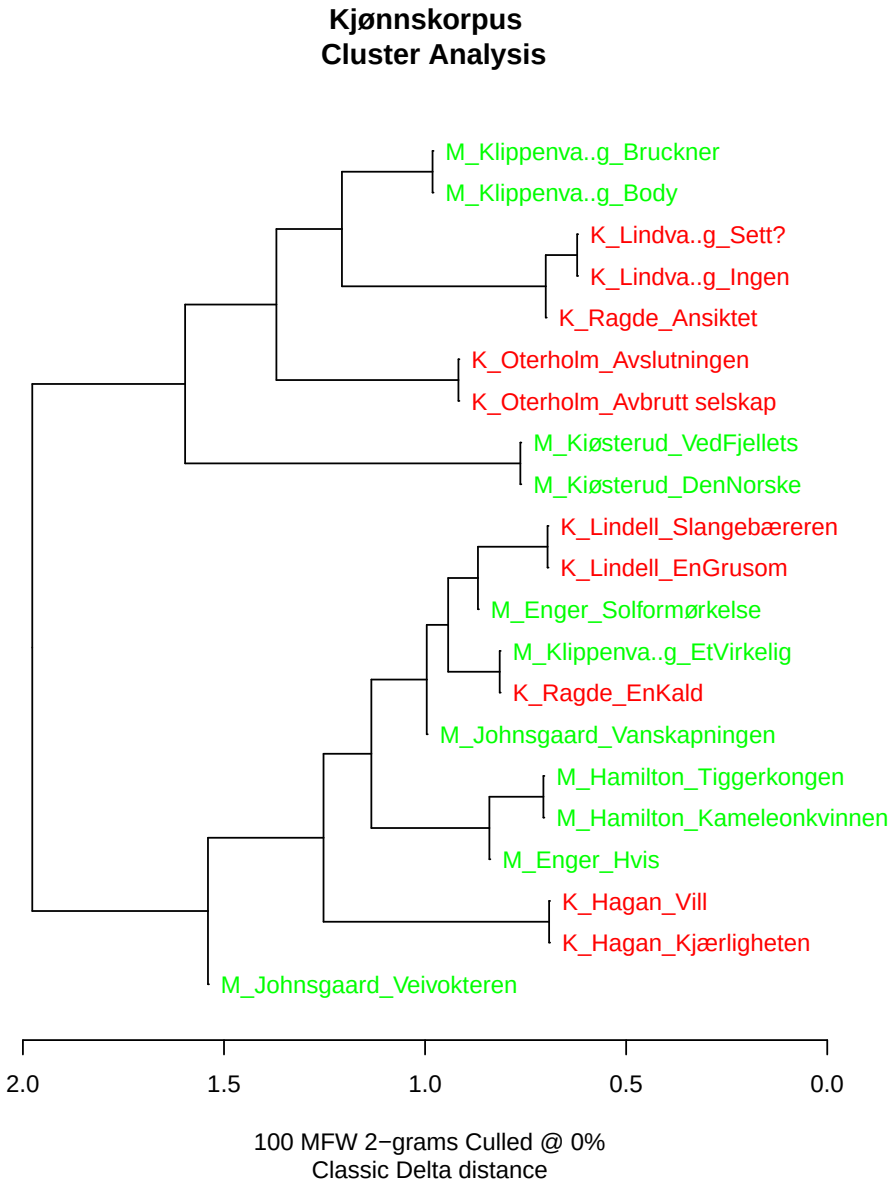
Grafer fra forsøk i “Kjønnskorpuset”



Figur E.1: Med n-gram (n=3) av bokstaver og MFW=100



Figur E.2: Med n-gram (n=3) av bokstaver og MFW=0



Figur E.3: Med n-gram (n=2) av ord MFW=100

Tillegg F

Programmet “merit.r”

```
# script for Victoria Troland
# Skrevet av Prof. Koenraad De Smedt, modfisert av Victoria Troland
# Vår 2014
# Programmet skriver .csv-tabell om til .tex-tabell og lager søylediagram
# av IG tallene og printer de ut i .pdf-fil.

require (xtable)
# if needed: set working directory to where files are
setwd("/Users/victoriatroland/Desktop/INESS korpus/InfoGain/")
# read table with merit etc. produced by Weka
morig <- read.table("3IGfrek.txt", header=FALSE)
# make new data frame with selected columns, row names and column names
mall <- data.frame(morig[1],morig[3],morig[4],morig[6])
colnames(mall) <- c("meritt","+- meritt", "rang", "+- rang")
rownames(mall) <- t(morig[8])

#take next table, glue together
morig2 <- read.table("3IGkvant.txt", header=FALSE)
# make new data frame with selected columns, row names and column names
mall2 <- data.frame(morig2[1],morig2[3],morig2[4],morig2[6])
colnames(mall2) <- c("meritt","+- meritt", "rang", "+- rang")
rownames(mall2) <- t(morig2[8])

# Add sort and add tables together
morig <- morig[order(morig$V8),]
morig2 <- morig2[order(morig2$V8),]

morig3 <- data.frame((morig[1]+morig2[1]),(morig[3]+morig2[3]),
```

```
(morig[4]+morig2[4]),(morig[6]+morig2[6]))
colnames(morig3) <- c("meritt","+- meritt", "rang", "+- rang")
rownames(morig3) <- t(morig[8])

morig4 <- data.frame(morig3,rownames(morig3))
morig4 <- morig4[order(-morig4$meritt),]

morig3 <- morig3[order(-morig3$meritt),]
print(xtable(morig3, caption="Attributtenes merit og rangering", l
abel="tab:merit"), file="merit.tex", type="latex") # , hline.after=c(-1,0,0)



m <- data.frame(morig3[1],morig4$rownames.morig3.)
# sort in ascending order based on first column
ms <- m[order(-m[1]),]
colnames(ms) <- c("avg merit","att")
rownames(ms) <- ms$att
# plot
pdf("Rplot-merit.pdf")
par(mar=c(4,6,1,1)+0.1) # clockwise from bottom
barplot(t(ms), horiz=TRUE, las=2, xlab=" Attributenes gjennomsnittlige merit")
dev.off()
# plot for morig1 og morig2
n <- data.frame(morig[1],morig[,8])
# sort in ascending order based on first column
ns <- n[order(-n[1]),]
colnames(ns) <- c("avg.merit","att")
rownames(ns) <- ns$att
# plot
pdf("Rplotfrek-merit.pdf")
par(mar=c(4,6,1,1)+0.1) # clockwise from bottom
barplot(t(ns), horiz=TRUE, las=2, xlab=" Attributenes gjennomsnittlige merit")
dev.off()
p <- data.frame(morig2[1],morig[,8])
# sort in ascending order based on first column
ps <- p[order(-p[1]),]
colnames(ps) <- c("avg.merit","att")
rownames(ps) <- ps$att
# plot
pdf("Rplotkvant-merit.pdf")
```

```
par(mar=c(4,6,1,1)+0.1) # clockwise from bottom
barplot(t(ps), horiz=TRUE, las=2, xlab=" Attributenes gjennomsnittlige merit")
dev.off()
```

Tillegg G

Ordliste - egne oversettelser fra engelsk til norsk

Engelsk	**Norsk**
stylometry	stilometri
cross-validation	kryssvalidering
10-fold	10-delt
cross-matrix	krysningstabell
authorship attribution	forfatterattribuering
authorship verification	forfatterverifisering
supervised	overvåket
unsupervised	ikke-overvåket
forensic linguistics	forensisk lingvistikk
shallow	grunt
filler words	funksjonsord
rewrite rules	omskrivningsregler
clusters	klynger

Tabell G.1: Egne oversettelser fra engelsk til norsk

Tillegg H

Programmet “korpusliste.r” og “inessfrekvenser.r”

H.1 “korpusliste.r”

```
# script for Victoria Troland
# Programmet skriver ut .csv-fil om til .tex-tabell.
# Skrevet av Prof. Koenraad De Smedt
# Vår 2015

require (xtable)
# set working directory to where files are
setwd("~/Documents/cursus/dasp350/Victoria Troland/")
# read csv file with corpus overview
kl <- read.table("korpusliste.csv", sep="\t", header=TRUE, row.names=1)
# print latex table with selected columns
print(xtable(kl[,c(1:3,7:9)], caption="Oversikt over innhold i korpuset",
label="table:korpusliste", align="rlp{36mm}crll"), file="korpusliste.tex",
type="latex", latex.environments="center") # vurder "sideways"
print(xtable(kl[,c(5:6,10:14)], caption="Egenskaper av innhold i korpuset",
label="table:korpusliste"), file="korpusliste2.tex", type="latex",
latex.environments="center")
sorted <- kl[order(-kl$Ord),]
ord1000 <- sorted[,6]/1000
pdf("ord.pdf")
barplot(ord1000, las=2, ylab="Ord (x 1000)", names.arg=rownames(sorted))
dev.off()
```

H.2 "inessfrekvenser.r"

```
# Skript for å lage latex tabell av .csv-fil
# Basert på programmet "korpusliste.r" av Prof. Koenraad De Smedt,
# modifisert av Victoria Troland

require (xtable)
# set working directory to where files are
setwd("~/Documents/Desktop/INESS korpus/R/")
# read csv file with corpus overview
kl <- read.table("cutfile.csv", sep=",", header=TRUE, row.names=1)
# print latex table with selected columns
print(xtable(kl[,c(1:2,5:9)], caption="Oversikt over frekvenser",
label="table:inessfrek", align="llp{25mm}llllll"), file="frek1.tex",
type="latex", latex.environments="center") # vurder "sideways"
print(xtable(kl[,c(10:16)], caption="Oversikt over frekvenser, del 2",
label="table:inessfrek2", align="llllllll"), file="frek2.tex", type="latex",
latex.environments="center")
print(xtable(kl[,c(17:22)], caption="Oversikt over frekvenser, del 3",
label="table:inessfrek3", align="lllllll"), file="frek3.tex", type="latex",
latex.environments="center")
print(xtable(kl[,c(25:30)], caption="Oversikt over frekvenser",
label="table:inessfrek4",
align="lllllll"), file="frek4.tex", type="latex", latex.environments="center")
```

Tillegg I

Liste over frekvenser hentet fra INESS[1]

[1] Laget med skriptet “inessfrekvenser.r”

	Forfatter	Tittel	KoordTot	KoordMen	KoordKomma	KoordEller	KoordOg
ar1	Ragde	En kald dag i helvete	2754	98	1042	42	1535
ar2	Ragde	Ansiktet som solen	2165	82	834	41	1161
mj1	Johnsgaard	Vanskapningen	1118	60	120	48	865
mj2	Johnsgaard	Veivokteren	1037	28	238	80	664
ph1	Hagan	Vill og vakker	3542	355	587	66	2375
ph2	Hagan	Kjærlighetens triumf	3178	279	601	69	2121
ul1	Lindell	Slangebæreren	4310	210	1558	151	2341
ul2	Lindell	En grusom kvinnes bekjennelser	2381	168	606	64	1479
ao1	Oterholm	Avbrutt selskap	207	6	59	20	115
ao2	Oterholm	Avslutningen	390	11	95	69	204
ok1	Klippenvåg	Body & Soul	1708	112	772	30	765
ok2	Klippenvåg	Bruckner, en lengsel	1068	56	561	16	410
ok3	Klippenvåg	Et virkelig liv	3899	245	2214	58	1292
el1	Lindvåg	Ingen kan nå meg	2511	112	946	110	1302
el2	Lindv?g	Sett: ?	4791	173	1881	209	2441
re1	Enger	Solformørkelse	2125	140	486	67	1393
re2	Enger	Hvs noen skulle være så vemmelige	1833	92	780	58	873
dh1	Hamilton	Kameleonkvinnen	1497	207	212	66	958
dh2	Hamilton	Tiggerkongen	1517	187	208	70	991
ek1	Kiøsterud	Ved fjellets fot	387	1	307	2	77
ek2	Kiøsterud	Den norske sangeren	770	26	429	13	296

Tabell I.1: Oversikt over frekvenser

	Passiv+	Passiv-	Adj.pred	Adj.attr	Intetkj.nn	Hunkj.nn	Hankj.nn
ar1	10	428	790	1652	5268	2429	5581
ar2	4	290	790	1153	3822	1726	4014
mj1	39	325	882	995	5099	1186	6147
mj2	23	254	605	767	4147	531	5151
ph1	50	649	1679	2404	8368	2730	9080
ph2	40	643	1514	1901	7562	2569	8530
ul1	62	819	1829	4161	10963	3590	13837
ul2	23	327	1037	1889	4872	1371	6238
ao1	27	109	708	291	3488	1360	2650
ao2	47	200	1140	529	5049	750	3026
ok1	55	224	623	1096	3586	808	5236
ok2	36	192	542	803	2547	264	2625
ok3	122	661	1536	2269	8033	1848	8727
el1	42	459	1097	1518	5922	2669	5621
el2	30	731	1478	2897	8156	3236	8540
re1	20	329	745	1320	4382	1023	5809
re2	16	230	727	1014	3582	1131	4513
dh1	3	349	574	915	3421	289	3665
dh2	3	395	608	997	3542	345	3801
ek1	0	72	121	344	531	72	674
ek2	3	186	219	666	1541	204	2238

Tabell I.2: Oversikt over frekvenser, del 2

	Akkustativ	Dativ	Genitiv	Nominativ	Kasus.Tot	Oblik
ar1	0	0	40	6633	15172	8499
ar2	0	0	73	4959	10750	5718
mj1	0	0	62	7894	15264	7308
mj2	2	0	77	5586	11582	5917
ph1	0	0	441	12067	25796	13288
ph2	0	0	334	10483	22349	11532
ul1	0	0	358	14649	31368	16361
ul2	0	0	137	7000	14953	7816
ao1	0	0	13	4734	8898	4151
ao2	0	0	43	6308	11676	5325
ok1	5	0	55	5897	11530	5573
ok2	0	0	81	4370	8193	3742
ok3	0	0	91	12153	24853	12609
el1	0	0	163	7789	16814	8862
el2	0	0	334	9915	23134	12885
re1	0	0	57	6601	13603	6945
re2	0	0	57	5278	11147	5812
dh1	0	0	105	4817	10132	5210
dh2	0	0	121	5103	10838	5614
ek1	0	0	14	592	1406	800
ek2	0	0	60	1852	4702	2790

Tabell I.3: Oversikt over frekvenser, del 3

	Referanse+	Referanse-	Singularis	Pluralis	Ubestemt	Bestemt
ar1	799	10577	17643	5328	1829	4670
ar2	574	7830	13274	3732	1502	3379
mj1	1108	11936	17404	4547	2637	3058
mj2	914	8649	12583	4588	1190	2562
ph1	1447	21224	34529	5795	1298	5928
ph2	1377	18219	29995	4913	2040	4911
ul1	1937	24365	42179	9075	3500	10440
ul2	951	11115	19640	4200	1688	5061
ao1	931	7231	11830	1603	845	1646
ao2	1343	10087	16035	2537	1187	1909
ok1	952	8576	14542	2772	959	3248
ok2	460	7138	11090	1777	697	1817
ok3	1629	18643	31832	5316	2381	6926
el1	1102	13586	22217	4901	2111	4205
el2	1213	16278	29054	9824	3151	7603
re1	797	10840	16374	4449	1743	3942
re2	686	8369	13985	2913	1471	3331
dh1	511	8147	12378	3455	1101	2796
dh2	519	8419	12946	3773	1121	2877
ek1	84	886	1801	842	170	715
ek2	226	2914	6010	1484	443	2293

Tabell I.4: Oversikt over frekvenser

Tillegg J

Liste over søkene i INESS

Bestemthet: ”-” og ”+”:

```
 #x >DEF "\-" & #x >(NTYPE NSYN) 'common' :: title = "(En kald.*|Ansikt.*
|Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|
Avslut.*|Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|
Hvis noen s.*|Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

 #x >DEF "\+" & #x >(NTYPE NSYN) 'common' :: title = "(En kald.*|Ansikt.*
|Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|
Avslut.*|Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|
Hvis noen s.*|Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Tall: ”singularis” og ”pluralis”

```
 #x >(NUM) 'sg' :: title ="(En kald.*|Ansikt.*
|Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|
Avslut.*|Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|
Hvis noen s.*|Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

 #x >(NUM) 'pl' :: title ="(En kald.*|Ansikt.*
|Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|
Avslut.*|Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|
Hvis noen s.*|Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Grammatisk kjønn: ”FEM”, ”MASC” og ”NEUT”:

```
 #x >(FEM) "\+" :: title = "(En kald.*|
Ansikt.*|Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*
|En gruso.*|Avbr.*|Avslut.*|Body.*|Bruck.*|Et vir.*|
Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|Kjærligheten.* triumf|
Tigge.*|Ved fj.*|Den n.*)"
```

```
 #x >(MASC) "\+" :: title = "(En kald.*|
Ansikt.*|Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*
|En gruso.*|Avbr.*|Avslut.*|Body.*|Bruck.*|Et vir.*|
Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|Kjærligheten.* triumf|
Tigge.*|Ved fj.*|Den n.*)"

 #x >(NEUT) "\+" :: title = "(En kald.*|
Ansikt.*|Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*
|En gruso.*|Avbr.*|Avslut.*|Body.*|Bruck.*|Et vir.*|
Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|Kjærligheten.* triumf|
Tigge.*|Ved fj.*|Den n.*)"
```

Passiv: ”-” og ”+”:

```
#x >(PASSIVE) "\-" :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|
Avslut.*|Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

#x >(PASSIVE) "\+" :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|
Avslut.*|Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Referanse: ”-” og ”+”:

```
 #x >REF "\-" :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|
Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

 #x >REF "\+" :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|
Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Adjektiv type, predikativ og attributativ:

```
 #x >(ATYPE) 'predicative' :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|
Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

```
 #x >(ATYPE) 'attributative' :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|
Body.*|Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Koordingering, men, eller, og, komma og totalt:

```
 #x >(COORD-FORM) 'og' :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

#x >(COORD-FORM) 'men' :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

#x >(COORD-FORM) 'eller' :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

#x >(COORD-FORM) 'komma' :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

#x >(COORD-FORM) '.*' :: title = "(En kald.*|Ansikt.*|
Vansk.*|Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Kasus: nominativ, dativ, akkustativ, oblik, totalt og genitiv

```
#x >(CASE) 'obl' :: title = "(En kald.*|Ansikt.*|Vansk.*|
 Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
 Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
 Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

#x >(CASE) 'nom' :: title = "(En kald.*|Ansikt.*|Vansk.*|
 Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
 Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
 Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

```
#x >(CASE) 'gen' :: title = "(En kald.*|Ansikt.*|Vansk.*|
 Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
 Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
 Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

#x >(CASE) 'dat' :: title = "(En kald.*|Ansikt.*|Vansk.*|
 Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
 Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
 Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

#x >(CASE) 'acc' :: title = "(En kald.*|Ansikt.*|Vansk.*|
 Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
 Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
 Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"

#x >(CASE) '.*' :: title = "(En kald.*|Ansikt.*|Vansk.*|
 Veiv.*|.*vakker|Kameleon.*|Slang.*|En gruso.*|Avbr.*|Avslut.*|Body.*|
 Bruck.*|Et vir.*|Ingen kan.*|Sett:.*|Solfor.*|Hvis noen s.*|
 Kjærligheten.* triumf|Tigge.*|Ved fj.*|Den n.*)"
```

Tillegg K

Oppsummering av resultatene til INESS-forsøkene av WEKA

K.1 Forsøk med kontinuerlige verdier

Baseline med kontinuerlige verdier

```
 === Evaluation on training set ===
=== Summary ===

Correctly Classified Instances           3               14.2857 %
Incorrectly Classified Instances        18               85.7143 %
Kappa statistic                          0
Mean absolute error                      0.1797
Root mean squared error                  0.2997
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.5      Johnsgaard.
                 0         0          0         0         0          0.5      Lindell.
                 0         0          0         0         0          0.5      Enger.
                 0         0          0         0         0          0.5      Ragde.
                 0         0          0         0         0          0.5      Oterholm.
                 0         0          0         0         0          0.5      Hagan.
                 0         0          0         0         0          0.5      Kiøsterud.
                 0         0          0         0         0          0.5      Hamilton.
```

```
               1         1         0.143    1         0.25      0.5      Klippenvåg.
                 0         0         0         0         0         0.5      Lindv?g
Weighted Avg.    0.143     0.143     0.02      0.143     0.036     0.5

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 0 0 0 0 0 0 0 2 0 | a = Johnsgaard.
 0 0 0 0 0 0 0 0 2 0 | b = Lindell.
 0 0 0 0 0 0 0 0 2 0 | c = Enger.
 0 0 0 0 0 0 0 0 2 0 | d = Ragde.
 0 0 0 0 0 0 0 0 2 0 | e = Oterholm.
 0 0 0 0 0 0 0 0 2 0 | f = Hagan.
 0 0 0 0 0 0 0 0 2 0 | g = Kiøsterud.
 0 0 0 0 0 0 0 0 2 0 | h = Hamilton.
 0 0 0 0 0 0 0 0 3 0 | i = Klippenvåg.
 0 0 0 0 0 0 0 0 2 0 | j = Lindv?g
```

FT med kontinuerlige verdier

```
 === Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           0               0      %
Incorrectly Classified Instances        21             100      %
Kappa statistic                         -0.1484
Mean absolute error                      0.2
Root mean squared error                  0.4472
Relative absolute error                108.0198 %
Root relative squared error            144.757  %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0.316     0         0         0         0.342    Johnsgaard.
                   0         0         0         0         0         0.5      Lindell.
                   0         0         0         0         0         0.5      Enger.
                   0         0         0         0         0         0.5      Ragde.
                  0         0         0         0         0         0.5      Oterholm.
```

```
                 0        0         0        0        0         0.5      Hagan.
                 0        0         0        0        0         0.5      Ki?sterud.
                 0        0         0        0        0         0.5      Hamilton.
                 0        0.833     0        0        0         0.083    Klippenv?g.
                 0        0         0        0        0         0.5      Lindv?g.
Weighted Avg.    0        0.149     0        0        0         0.425

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 0 0 0 0 0 0 0 2 0 | a = Johnsgaard.
 0 0 0 0 0 0 0 0 2 0 | b = Lindell.
 1 0 0 0 0 0 0 0 1 0 | c = Enger.
 2 0 0 0 0 0 0 0 0 0 | d = Ragde.
 0 0 0 0 0 0 0 0 2 0 | e = Oterholm.
 0 0 0 0 0 0 0 0 2 0 | f = Hagan.
 0 0 0 0 0 0 0 0 2 0 | g = Ki?sterud.
 0 0 0 0 0 0 0 0 2 0 | h = Hamilton.
 3 0 0 0 0 0 0 0 0 0 | i = Klippenv?g.
 0 0 0 0 0 0 0 0 2 0 | j = Lindv?g.
```

NBM med kontinuerlige verdier

```
 === Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           1               4.7619 %
Incorrectly Classified Instances        20              95.2381 %
Kappa statistic                         -0.0909
Mean absolute error                      0.1786
Root mean squared error                  0.311
Relative absolute error                 96.4394 %
Root relative squared error            100.6801 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0        0         0        0        0         0.184    Johnsgaard.
                 0        0         0        0        0         0.105    Lindell.
```

```
                 0         0         0         0         0         0.105    Enger.
                 0         0         0         0         0         0.158    Ragde.
                 0         0         0         0         0         0.947    Oterholm.
                 0         0.316     0         0         0         0.132    Hagan.
                 0.5       0         1         0.5       0.667     0.526    Ki?sterud.
                 0         0         0         0         0         0.053    Hamilton.
                 0         0.778     0         0         0         0.259    Klippenv?g.
                 0         0         0         0         0         0.132    Lindv?g.
Weighted Avg.    0.048     0.141     0.095     0.048     0.063     0.26

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 0 0 0 0 0 0 0 2 0 | a = Johnsgaard.
 0 0 0 0 0 0 0 0 2 0 | b = Lindell.
 0 0 0 0 0 1 0 0 1 0 | c = Enger.
 0 0 0 0 0 2 0 0 0 0 | d = Ragde.
 0 0 0 0 0 0 0 0 2 0 | e = Oterholm.
 0 0 0 0 0 0 0 0 2 0 | f = Hagan.
 0 0 0 0 0 0 1 0 1 0 | g = Ki?sterud.
 0 0 0 0 0 0 0 0 2 0 | h = Hamilton.
 0 0 0 0 0 3 0 0 0 0 | i = Klippenv?g.
 0 0 0 0 0 0 0 0 2 0 | j = Lindv?g.
```

SVM med kontinuerlige verdier

```
 === Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          14               66.6667 %
Incorrectly Classified Instances         7               33.3333 %
Kappa statistic                          0.6297
Mean absolute error                      0.1628
Root mean squared error                  0.2771
Relative absolute error                 87.9018 %
Root relative squared error             89.7065 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===
```

```
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.974    Johnsgaard.
                 1         0.053      0.667     1         0.8        0.974    Lindell.
                 1         0.158      0.4       1         0.571      0.921    Enger.
                 1         0.158      0.4       1         0.571      0.947    Ragde.
                 1         0          1         1         1          1        Oterholm.
                 0.5       0          1         0.5       0.667      0.75     Hagan.
                 0.5       0          1         0.5       0.667      1        Ki?sterud.
                 1         0          1         1         1          1        Hamilton.
                 0.667     0          1         0.667     0.8        1        Klippenv?g.
                 0         0          0         0         0          0.789    Lindv?g.
Weighted Avg.    0.667     0.035      0.663     0.667     0.617      0.939

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 0 2 0 0 0 0 0 0 0 | a = Johnsgaard.
 0 2 0 0 0 0 0 0 0 0 | b = Lindell.
 0 0 2 0 0 0 0 0 0 0 | c = Enger.
 0 0 0 2 0 0 0 0 0 0 | d = Ragde.
 0 0 0 0 2 0 0 0 0 0 | e = Oterholm.
 0 1 0 0 0 1 0 0 0 0 | f = Hagan.
 0 0 0 1 0 0 1 0 0 0 | g = Ki?sterud.
 0 0 0 0 0 0 0 2 0 0 | h = Hamilton.
 0 0 1 0 0 0 0 0 2 0 | i = Klippenv?g.
 0 0 0 2 0 0 0 0 0 0 | j = Lindv?g.
```

k-NN med kontinuerlige verdier

```
 === Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          13               61.9048 %
Incorrectly Classified Instances         8               38.0952 %
Kappa statistic                          0.5768
Mean absolute error                      0.0762
Root mean squared error                  0.276
Relative absolute error                 41.1504 %
Root relative squared error             89.3459 %
Total Number of Instances               21
```

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.5      Johnsgaard.
                 1         0.158      0.4       1         0.571      0.921    Lindell.
                 0         0.158      0         0         0          0.421    Enger.
                 1         0.105      0.5       1         0.667      0.947    Ragde.
                 1         0          1         1         1          1        Oterholm.
                 0.5       0          1         0.5       0.667      0.75     Hagan.
                 1         0          1         1         1          1        Ki?sterud.
                 1         0          1         1         1          1        Hamilton.
                 0.667     0          1         0.667     0.8        0.833    Klippenv?g.
                 0         0          0         0         0          0.5      Lindv?g.
Weighted Avg.    0.619     0.04       0.61      0.619     0.581      0.789

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 0 2 0 0 0 0 0 0 0 | a = Johnsgaard.
 0 2 0 0 0 0 0 0 0 0 | b = Lindell.
 0 2 0 0 0 0 0 0 0 0 | c = Enger.
 0 0 0 2 0 0 0 0 0 0 | d = Ragde.
 0 0 0 0 2 0 0 0 0 0 | e = Oterholm.
 0 1 0 0 0 1 0 0 0 0 | f = Hagan.
 0 0 0 0 0 0 2 0 0 0 | g = Ki?sterud.
 0 0 0 0 0 0 0 2 0 0 | h = Hamilton.
 0 0 1 0 0 0 0 0 2 0 | i = Klippenv?g.
 0 0 0 2 0 0 0 0 0 0 | j = Lindv?g.
```

LMT med kontinuerlige verdier

```
  === Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           8               38.0952 %
Incorrectly Classified Instances        13               61.9048 %
Kappa statistic                          0.3106
Mean absolute error                      0.1193
Root mean squared error                  0.3064
```

```
Relative absolute error                 64.4386 %
Root relative squared error             99.1699 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.763    Johnsgaard.
                 0.5       0.158      0.25      0.5       0.333      0.842    Lindell.
                 0         0.105      0         0         0          0.763    Enger.
                 0         0.105      0         0         0          0.684    Ragde.
                 0.5       0          1         0.5       0.667      0.842    Oterholm.
                 0         0.053      0         0         0          0.711    Hagan.
                 0.5       0          1         0.5       0.667      0.632    Ki?sterud.
                 1         0.053      0.667     1         0.8        1        Hamilton.
                 1         0          1         1         1          1        Klippenv?g.
                 0         0.211      0         0         0          0.711    Lindv?g.
Weighted Avg.    0.381     0.065      0.421     0.381     0.378      0.805

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 0 1 0 0 0 0 0 0 1 | a = Johnsgaard.
 0 1 0 0 0 0 0 0 0 1 | b = Lindell.
 0 2 0 0 0 0 0 0 0 0 | c = Enger.
 0 0 1 0 0 0 0 0 0 1 | d = Ragde.
 0 0 0 0 1 0 0 1 0 0 | e = Oterholm.
 0 1 0 0 0 0 0 0 0 1 | f = Hagan.
 0 0 0 1 0 0 1 0 0 0 | g = Ki?sterud.
 0 0 0 0 0 0 0 2 0 0 | h = Hamilton.
 0 0 0 0 0 0 0 0 3 0 | i = Klippenv?g.
 0 0 0 1 0 1 0 0 0 0 | j = Lindv?g.
```

LADTree med kontinuerlige verdier

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          11               52.381  %
Incorrectly Classified Instances        10               47.619  %
```

```
Kappa statistic                          0.471
Mean absolute error                      0.0989
Root mean squared error                  0.2744
Relative absolute error                 53.3957 %
Root relative squared error             88.8353 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0        0.053     0          0        0          0.868    Johnsgaard.
                 0.5      0.053     0.5        0.5      0.5        0.789    Lindell.
                 0        0.158     0          0        0          0.579    Enger.
                 0.5      0.105     0.333      0.5      0.4        0.947    Ragde.
                 1        0.105     0.5        1        0.667      0.947    Oterholm.
                 0.5      0         1          0.5      0.667      0.895    Hagan.
                 0.5      0.053     0.5        0.5      0.5        0.947    Ki?sterud.
                 1        0         1          1        1          1        Hamilton.
                 0.667    0         1          0.667    0.8        0.981    Klippenv?g.
                 0.5      0         1          0.5      0.667      0.737    Lindv?g.
Weighted Avg.    0.524    0.05      0.603      0.524    0.533      0.875

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 0 1 0 1 0 0 0 0 0 | a = Johnsgaard.
 0 1 1 0 0 0 0 0 0 0 | b = Lindell.
 1 0 0 1 0 0 0 0 0 0 | c = Enger.
 0 0 0 1 0 0 1 0 0 0 | d = Ragde.
 0 0 0 0 2 0 0 0 0 0 | e = Oterholm.
 0 1 0 0 0 1 0 0 0 0 | f = Hagan.
 0 0 0 1 0 0 1 0 0 0 | g = Ki?sterud.
 0 0 0 0 0 0 0 2 0 0 | h = Hamilton.
 0 0 1 0 0 0 0 0 2 0 | i = Klippenv?g.
 0 0 0 0 1 0 0 0 0 1 | j = Lindv?g.
```

K.2 Forsøk med diskrete verdier

FT med diskrete verdier

```
 === Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           0               0      %
Incorrectly Classified Instances        21             100      %
Kappa statistic                         -0.1455
Mean absolute error                      0.2
Root mean squared error                  0.4472
Relative absolute error                108.0266 %
Root relative squared error            144.7763 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0.211      0         0         0          0.395    Enger
                 0         0.158      0         0         0          0.421    Lindv?g
                 0         0          0         0         0          0.5      Johnsgaard
                 0         0.778      0         0         0          0.111    Klippenvåg
                 0         0          0         0         0          0.5      Hagan
                 0         0          0         0         0          0.5      Ragde
                 0         0          0         0         0          0.5      Oterholm
                 0         0          0         0         0          0.5      Lindell
                 0         0          0         0         0          0.5      Kiøsterud
                 0         0          0         0         0          0.5      Hamilton
Weighted Avg.    0          0.146      0          0          0           0.427

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 1 0 1 0 0 0 0 0 0 | a = Enger
 0 0 0 2 0 0 0 0 0 0 | b = Lindv?g
 0 0 0 2 0 0 0 0 0 0 | c = Johnsgaard
 2 1 0 0 0 0 0 0 0 0 | d = Klippenvåg
 2 0 0 0 0 0 0 0 0 0 | e = Hagan
 0 0 0 2 0 0 0 0 0 0 | f = Ragde
 0 1 0 1 0 0 0 0 0 0 | g = Oterholm
 0 0 0 2 0 0 0 0 0 0 | h = Lindell
 0 0 0 2 0 0 0 0 0 0 | i = Kiøsterud
 0 0 0 2 0 0 0 0 0 0 | j = Hamilton
```

NBM med diskrete verdier

```
 === Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          16               76.1905 %
Incorrectly Classified Instances         5               23.8095 %
Kappa statistic                          0.7348
Mean absolute error                      0.0734
Root mean squared error                  0.1835
Relative absolute error                 39.6259 %
Root relative squared error             59.4115 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0.053      0         0         0          0.737    Enger
                 0.5       0.053      0.5       0.5       0.5        0.895    Lindv?g
                 1         0          1         1         1          1        Johnsgaard
                 1         0          1         1         1          1        Klippenvåg
                 1         0          1         1         1          1        Hagan
                 0.5       0.053      0.5       0.5       0.5        0.921    Ragde
                 1         0          1         1         1          1        Oterholm
                 0.5       0.053      0.5       0.5       0.5        0.921    Lindell
                 1         0          1         1         1          1        Kiøsterud
                 1         0.053      0.667     1         0.8        1        Hamilton
Weighted Avg.    0.762     0.025      0.73      0.762     0.743      0.95

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 0 0 0 0 0 0 1 0 1 | a = Enger
 0 1 0 0 0 1 0 0 0 0 | b = Lindv?g
 0 0 2 0 0 0 0 0 0 0 | c = Johnsgaard
 0 0 0 3 0 0 0 0 0 0 | d = Klippenvåg
 0 0 0 0 2 0 0 0 0 0 | e = Hagan
 0 1 0 0 0 1 0 0 0 0 | f = Ragde
 0 0 0 0 0 0 2 0 0 0 | g = Oterholm
```

```
 1 0 0 0 0 0 0 1 0 0 | h = Lindell
 0 0 0 0 0 0 0 0 2 0 | i = Kiøsterud
 0 0 0 0 0 0 0 0 0 2 | j = Hamilton
```

SVM med diskrete verdier

```
 === Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          14               66.6667 %
Incorrectly Classified Instances         7               33.3333 %
Kappa statistic                          0.6288
Mean absolute error                      0.1623
Root mean squared error                  0.2763
Relative absolute error                 87.6787 %
Root relative squared error             89.4406 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0.105      0         0         0          0.316    Enger
                 0         0.053      0         0         0          0.895    Lindv?g
                 1         0          1         1         1          1        Johnsgaard
                 1         0          1         1         1          1        Klippenvåg
                 1         0          1         1         1          1        Hagan
                 0.5       0.105      0.333     0.5       0.4        0.921    Ragde
                 1         0          1         1         1          1        Oterholm
                 0         0.105      0         0         0          0.789    Lindell
                 1         0          1         1         1          1        Kiøsterud
                 1         0          1         1         1          1        Hamilton
Weighted Avg.    0.667     0.035      0.651     0.667     0.657      0.897

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 0 0 0 0 0 0 2 0 0 | a = Enger
 0 0 0 0 0 0 2 0 0 0 | b = Lindv?g
 0 0 2 0 0 0 0 0 0 0 | c = Johnsgaard
 0 0 0 3 0 0 0 0 0 0 | d = Klippenvåg
```

```
 0 0 0 0 2 0 0 0 0 0 | e = Hagan
 0 1 0 0 0 1 0 0 0 0 | f = Ragde
 0 0 0 0 0 0 2 0 0 0 | g = Oterholm
 2 0 0 0 0 0 0 0 0 0 | h = Lindell
 0 0 0 0 0 0 0 0 2 0 | i = Kiøsterud
 0 0 0 0 0 0 0 0 0 2 | j = Hamilton
```

k-NN med diskrete verdier

```
 === Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          14               66.6667 %
Incorrectly Classified Instances         7               33.3333 %
Kappa statistic                          0.6288
Mean absolute error                      0.0667
Root mean squared error                  0.2582
Relative absolute error                 36.0089 %
Root relative squared error             83.5867 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0.105      0         0         0          0.447    Enger
                 0         0.053      0         0         0          0.474    Lindv?g
                 1         0          1         1         1          1        Johnsgaard
                 1         0          1         1         1          1        Klippenvåg
                 1         0          1         1         1          1        Hagan
                 0.5       0.105      0.333     0.5       0.4        0.697    Ragde
                 1         0          1         1         1          1        Oterholm
                 0         0.105      0         0         0          0.447    Lindell
                 1         0          1         1         1          1        Kiøsterud
                 1         0          1         1         1          1        Hamilton
Weighted Avg.    0.667     0.035      0.651     0.667     0.657      0.816

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 0 0 0 0 0 0 2 0 0 | a = Enger
```

```
 0 0 0 0 0 2 0 0 0 0 | b = Lindv?g
 0 0 2 0 0 0 0 0 0 0 | c = Johnsgaard
 0 0 0 3 0 0 0 0 0 0 | d = Klippenvåg
 0 0 0 0 2 0 0 0 0 0 | e = Hagan
 0 1 0 0 0 1 0 0 0 0 | f = Ragde
 0 0 0 0 0 0 2 0 0 0 | g = Oterholm
 2 0 0 0 0 0 0 0 0 0 | h = Lindell
 0 0 0 0 0 0 0 0 2 0 | i = Kiøsterud
 0 0 0 0 0 0 0 0 0 2 | j = Hamilton
```

LMT med diskrete verdier

```
 === Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          13               61.9048 %
Incorrectly Classified Instances         8               38.0952 %
Kappa statistic                          0.5758
Mean absolute error                      0.0878
Root mean squared error                  0.2386
Relative absolute error                 47.4184 %
Root relative squared error             77.2571 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0.105      0         0         0          0.421    Enger
                 0         0.105      0         0         0          0.421    Lindv?g
                 1         0          1         1         1          1        Johnsgaard
                 1         0          1         1         1          1        Klippenvåg
                 1         0          1         1         1          1        Hagan
                 0         0.053      0         0         0          0.842    Ragde
                 1         0.105      0.5       1         0.667      0.947    Oterholm
                 0         0.053      0         0         0          0.921    Lindell
                 1         0          1         1         1          1        Kiøsterud
                 1         0          1         1         1          1        Hamilton
Weighted Avg.    0.619     0.04       0.571     0.619     0.587      0.862

=== Confusion Matrix ===
```

```
 a b c d e f g h i j   <-- classified as
 0 1 0 0 0 0 0 1 0 0 | a = Enger
 0 0 0 0 0 1 1 0 0 0 | b = Lindv?g
 0 0 2 0 0 0 0 0 0 0 | c = Johnsgaard
 0 0 0 3 0 0 0 0 0 0 | d = Klippenvåg
 0 0 0 0 2 0 0 0 0 0 | e = Hagan
 0 1 0 0 0 0 1 0 0 0 | f = Ragde
 0 0 0 0 0 0 2 0 0 0 | g = Oterholm
 2 0 0 0 0 0 0 0 0 0 | h = Lindell
 0 0 0 0 0 0 0 0 2 0 | i = Kiøsterud
 0 0 0 0 0 0 0 0 0 2 | j = Hamilton
```

LADTree med diskrete verdier

```
 === Stratified cross-validation ===
 === Summary ===

Correctly Classified Instances           8               38.0952 %
Incorrectly Classified Instances        13               61.9048 %
Kappa statistic                          0.3071
Mean absolute error                      0.1339
Root mean squared error                  0.3216
Relative absolute error                 72.3349 %
Root relative squared error            104.1031 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0.158     0         0         0          0.447    Enger
                 1         0.053     0.667     1         0.8        0.974    Lindv?g
                 0         0.211     0         0         0          0.763    Johnsgaard
                 1         0.111     0.6       1         0.75       0.963    Klippenvåg
                 0         0         0         0         0          0.684    Hagan
                 0         0         0         0         0          0.711    Ragde
                 0.5       0         1         0.5       0.667      0.974    Oterholm
                 0         0.105     0         0         0          0.658    Lindell
                 0.5       0.053     0.5       0.5       0.5        0.842    Kiøsterud
                 0.5       0         1         0.5       0.667      0.974    Hamilton
```

```
Weighted Avg.    0.381     0.071      0.387     0.381     0.358      0.807

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 0 0 1 1 0 0 0 0 0 0 | a = Enger
 0 2 0 0 0 0 0 0 0 0 | b = Lindv?g
 1 0 0 1 0 0 0 0 0 0 | c = Johnsgaard
 0 0 0 3 0 0 0 0 0 0 | d = Klippenvåg
 0 0 1 0 0 0 0 1 0 0 | e = Hagan
 0 1 0 0 0 0 0 0 1 0 | f = Ragde
 0 0 1 0 0 0 1 0 0 0 | g = Oterholm
 2 0 0 0 0 0 0 0 0 0 | h = Lindell
 0 0 0 0 0 0 0 1 1 0 | i = Kiøsterud
 0 0 1 0 0 0 0 0 0 1 | j = Hamilton
```

K.3 Information Gain

Information Gain med kontinuerlige verdier

```
 === Run information ===

Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     wekafreq-weka.filters.unsupervised.attribute.Remove-R17
Instances:    21
Attributes:   17
              CoordTot
              CoordMen
              CoordKomma
              CoordEller
              CoordOg
              Passiv
              Referanse
              AdjType
              Hunkjoenn
              Hankjoenn
              Intetkjoenn
              Genitiv
```

```
              Nominativ
              Oblik
              Tall
              Bestemthet
              Forfatter
Evaluation mode:10-fold cross-validation

=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 2.093 +- 0.21      3   +- 1.18    5 CoordOg
 2.073 +- 0.122     3.6 +- 1.28    2 CoordMen
 2.026 +- 0.308     4.4 +- 2.84    3 CoordKomma
 1.943 +- 0.351     5.3 +- 2.72   13 Nominativ
 1.978 +- 0.492     5.5 +- 3.77    6 Passiv
 1.922 +- 0.25      5.6 +- 2.42    4 CoordEller
 1.905 +- 0.37      5.7 +- 3.55   14 Oblik
 1.509 +- 0.27      9.1 +- 1.51   10 Hankjoenn
 1.253 +- 0.877     9.2 +- 5.06    1 CoordTot
 1.466 +- 0.332     9.7 +- 2.19   16 Bestemthet
 1.327 +- 0.261    10.4 +- 2.15   15 Tall
 1.216 +- 0.262    11.9 +- 1.87    9 Hunkjoenn
 0.807 +- 0.765    12.4 +- 3.77    8 AdjType
 0.693 +- 0.761    12.5 +- 4.13   11 Intetkjoenn
 1.027 +- 0.24     12.8 +- 1.08   12 Genitiv
 0.116 +- 0.349    14.9 +- 0.94    7 Referanse
```

Information Gain med diskrete verdier

```
 === Run information ===

Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     wekaquant-weka.filters.unsupervised.attribute.Remove-R17
Instances:    21
Attributes:   17
              CoordTot
              CoordMen
              CoordKomma
              CoordEller
```

```
               CoordOg
               Passiv
               Referanse
               AdjType
               Hunkjoenn
               Hankjoenn
               Intetkjoenn
               Genitiv
               Nominativ
               Oblik
               Tall
               Bestemthet
               Forfatter
Evaluation mode:10-fold cross-validation

=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 1.818 +- 0.05      1.3 +- 0.64    3 CoordKomma
 1.529 +- 0.197     4.2 +- 1.66    4 CoordEller
 1.508 +- 0.033     4.2 +- 1.54    6 Passiv
 1.496 +- 0.138     4.7 +- 1.68   13 Nominativ
 1.37  +- 0.333     5.6 +- 3.56   14 Oblik
 1.144 +- 0.604     7.1 +- 4.41    9 Hunkjoenn
 1.015 +- 0.234     8.6 +- 2.54   10 Hankjoenn
 0.787 +- 0.8       8.8 +- 5.1     2 CoordMen
 0.921 +- 0.398    10.1 +- 2.55    5 CoordOg
 0.853 +- 0.404    10.3 +- 3.74   11 Intetkjoenn
 0.815 +- 0.471    10.4 +- 2.5     8 AdjType
 0.861 +- 0.472    10.5 +- 2.58   15 Tall
 0.512 +- 0.534    11.7 +- 2.76   16 Bestemthet
 0.642 +- 0.628    11.9 +- 4.66    1 CoordTot
 0.417 +- 0.418    12.4 +- 2.54   12 Genitiv
 0     +- 0        14.2 +- 1.17    7 Referanse
```